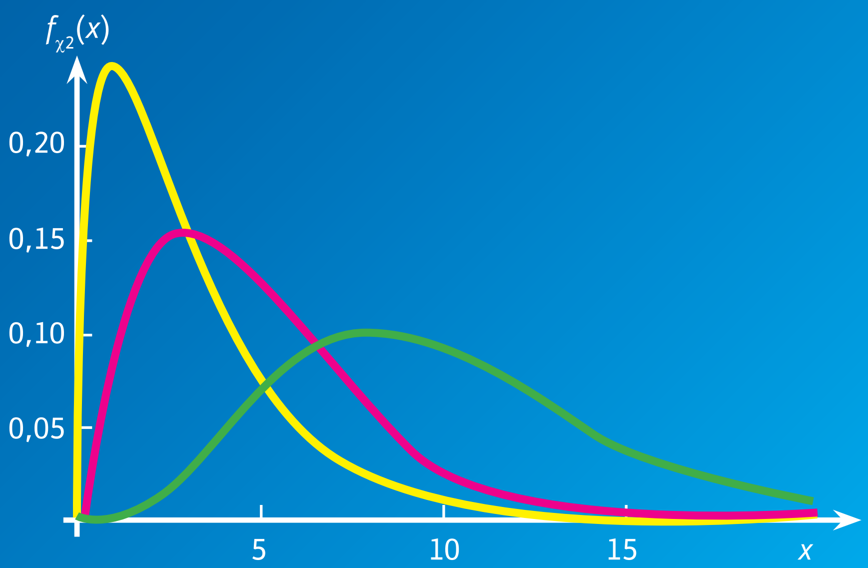
Уральский федеральный университет
имени первого Президента России Б. Н. Ельцина
Институт фундаментального образования

И. В. ГРЕБЕННИКОВА

МЕТОДЫ МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Учебно-методическое пособие



Министерство образования и науки Российской Федерации
Уральский федеральный университет
имени первого Президента России Б. Н. Ельцина

И. В. Гребенникова

МЕТОДЫ МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Учебно-методическое пособие

Рекомендовано методическим советом УрФУ
для студентов, обучающихся по направлениям подготовки
09.03.02 «Информационные системы и технологии»
и 39.04.04 «Организация работы с молодежью»

Екатеринбург
Издательство Уральского университета
2015

УДК 519.6:001.891(075.8)
ББК 22.171я73+72в6я73
Г79

Рецензенты:
кафедра «Информационных технологий» ГАОУ ДПО Института развития образования (завкафедрой д-р пед. наук, проф. Л. И. Долинер); завлабораторией рентгеновской спектроскопии д-р физ.-мат. наук М. А. Коротин (Институт физики металлов УрО РАН).

Гребенникова, И. В.
Г79 Методы математической обработки экспериментальных данных : учебно-методическое пособие / И. В. Гребенникова. — Екатеринбург : Изд-во Урал. ун-та, 2015. — 124 с.

ISBN 978-5-7996-1456-0

Представлены основные разделы курса «Методы математической обработки экспериментальных данных» технического вуза. Каждый раздел содержит теоретическую часть, методику проведения лабораторных работ, варианты заданий для выполнения самостоятельных работ.

Предназначено для студентов всех специальностей и форм обучения.

Библиогр.: 9 назв. Рис. 41. Прил. 4.

УДК 519.6:001.891(075.8)
ББК 22.171я73+72в6я73

ISBN 978-5-7996-1456-0

Оглавление

Предисловие

Каждая глава пособия состоит из трех частей: теоретической, методической и практической. В первой части содержится теоретический материал справочного характера: понятия, определения, утверждения, формулы по курсу «Математические методы обработки экспериментальных данных», а также примеры решения задач, графические иллюстрации. Во второй части приведена методика проведения лабораторных работ. Третья часть включает варианты заданий для выполнения самостоятельных работ.

По содержанию данное пособие соответствует требованиям ФГОС ВПО для специальностей технических вузов и включает в себя в соответствии с учебной программой основные разделы:

- построение вариационных рядов статистических распределений и расчет числовых характеристик;
- статистическое оценивание параметров распределения;
- построение математических моделей парных линейной и нелинейной корреляций;
- построение эмпирических и теоретических кривых распределения с нормальным теоретическим в соответствии с критерием согласия.

Предназначено для студентов вузов всех форм обучения, может быть использовано преподавателями при организации и проведении лабораторных занятий по методам обработки экспериментальных данных.

Введение

Методы обработки экспериментальных данных начали разрабатываться более двух веков тому назад в связи с необходимостью решения практических задач по агробиологии, медицине, экономике, социологии. Полученные при этом результаты составили фундамент такой научной дисциплины, как математическая статистика. В научный оборот слово «статистика» ввел немецкий ученый, профессор философии и права Готфрид Ахенваль (1719—1772), определяя этим термином круг вопросов, относящихся к государствоведению: описание государства, его устройства, быта населения, естественных условиях климата и др. Близким к современному пониманию сущности статистки было направление политической арифметики, основоположниками которой являлись английские исследователи: Джон Граунт (1620—1674), торговец, член Лондонского королевского общества; Уильям Петти (1623—1687), экономист, врач, доктор физики, профессор анатомии, изобретатель копировальной машины. Граунт Д. на основе обработки данных о естественном движении населения сделал вывод о существовании тесной связи между интенсивностью демографических процессов и особенностями социальной жизни людей, открыл некоторые закономерности массовых общественных явлений, используя собственный метод обработки и анализа первичных данных. Он попытался также построить таблицы смертности (или таблицы «дожития», показывающие процент доживших до определенного возраста), используемые для целей страхования. Исследования Петти относятся к политической экономии. Наибольшую известность получили его работы «Эссе о политической арифметике» (1683 г.) и «Политический обзор, или Анатомия Ирландии» (1672 г.), в которых социально-экономические вопросы изучались на основе числовых данных. Петти У. показал, что можно реконструировать информацию при отсутствии достаточного объема исходных данных (то есть с помощью различных расчетов найти нужные характеристики).

Возникновению математической статистики как науки способствовало появление теории вероятностей, развитие ее методов, использование в практических приложениях. Становление статистики связано с именами выдающихся ученых, в числе которых Пьер Симон Лаплас (1749—1827), Симеон Дени Пуассон (1781—1840), Жан Батист Фурье (1768—1830), Ламбер Адольф

Кетле (1796–1874). Они заложили основы современной статистической методологии, активно применяли полученные методы для установления закономерностей в общественных явлениях.

Современный уровень естественнонаучного эксперимента характеризуется большими потоками информации. При этом визуальный просмотр данных, не говоря уже об анализе, невозможен без применения ЭВМ. Обработка результатов экспериментов предполагает знание основных понятий и методов теории вероятностей и математической статистики. Выявление характерных классов задач в обработке экспериментальных данных и стандартных методов их решения позволяет выделить обработку результатов экспериментов из многообразия задач прикладной статистики. Как правило, основным подходом в решении многих задач является метод наименьших квадратов (МНК) в его различных модификациях. Однако МНК эффективно работает только для линейных моделей, а на практике встречаются ситуации, когда связь искомого параметра с измеряемой величиной сугубо нелинейная. В этом случае применяют нелинейный МНК или другие методы обработки. Знакомство со всеми этими методами расширяет арсенал средств, находящихся в распоряжении обработчика, что особенно важно в сложных случаях, например, когда измерения производятся при воздействии большого числа факторов, мешающих их проведению.

Появление электронных таблиц (табличных процессоров) привело к тому, что статистические методы, ранее доступные лишь узкому кругу математиков, стали использоваться широким кругом специалистов разных областей.

Дальнейшее развитие программного обеспечения привело к созданию большого количества прикладных пакетов по статистике. Удобной универсальной вычислительной средой для решения задач обработки экспериментальных данных является табличный процессор Microsoft Excel.

1. Вариационные ряды и их характеристики

Статистика изучает методы сбора и анализа результатов наблюдений массовых случайных явлений в целях выявления существующих закономерностей. Типичная задача математической статистики — на основании результатов наблюдений оценить вероятность случайного события или характеристики случайной величины. При решении любой задачи математической статистики имеется два источника информации. Первый источник — результаты наблюдений (экспериментов), причем процесс наблюдений может корректироваться на основании предварительных результатов (так называемый последовательный анализ). Второй источник — априорная (доопытная) информация о свойствах изучаемого объекта, накопленная к текущему моменту. Эта информация отражается в статистической модели, выбираемой при решении задачи. Следует заметить, что степень обоснованности применения априорной информации зависит от компетентности и добросовестности конкретного исследователя и неверные исходные допущения могут существенно исказить результат статистического анализа.

1.1. Первичная обработка экспериментальных данных

Пусть все возможные значения случайной величины X, получаемые при экспериментальных измерениях, образуют множество $\{x_1, x_2, ...\}$, которое называется *генеральной совокупностью значений X.*

Выборочная совокупность, или выборка, случайной величины X — случайно выбранное подмножество генеральной совокупности $\{x_1, x_2, ..., x_n\}$, полученное при проведении в одинаковых условиях n независимых измерений случайной величины X, где n — ***объем выборки.***

Значения x_i располагают в порядке возрастания: $\{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$, $x^{(1)} \le x^{(2)} \le ... \le x^{(n)}$. Может оказаться, что некоторые *варианты* x_i в выборке встречаются несколько раз. Число n_i, показывающее сколько раз встречается варианта x_i в выборочной совокупности, называется *частотой* (*эмпирической частотой*) результата x_i. Частоты вариант называют их *весами*. Отношение $w_i = \dfrac{n_i}{n}$ частоты к объему выборки называют *относительной частотой* значения x_i. Очевидно, что $\sum\limits_i \dfrac{n_i}{n} = 1$.

Вариационный ряд (или *статистическое распределение*) — это упорядоченный ряд вариант с соответствующими им весами: (x_i, n_i).

Различают *дискретные* и *непрерывные* вариационные ряды. *Дискретный* ряд записывают так:

Варианты, x_i	x_1	x_2	...	x_k
Частоты, n_i	n_1	n_2	...	n_k

Здесь n_i — частота появления результата x_i.

Если объем выборки большой ($n > 30$), то результаты наблюдений сводят в *интервальный вариационный ряд*, который формируется следующим образом. Вычисляют *размах R варьирования* признака X как разность между наибольшим $x_{\max}$ и наименьшим $x_{\min}$ значениями признака:

$$R = x_{\max} - x_{\min}.$$

Размах R варьирования признака X делят на k равных частей и таким образом определяют число интервалов. Число k частичных интервалов выбирают, пользуясь одним из следующих правил:

1) $6 \le k \le 20$;
2) $k \approx \sqrt{n}$;
3) $k \approx 1 + \log_2 n \approx 1 + 3,221 \lg n$.

При небольшом объеме n выборки число k принимают равным от 6 до 10. Длина каждого частичного интервала определяется по формуле

$$h = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{k}.$$

Величину h обычно округляют до некоторого значения d. Например, если результаты x_i признака X — целые числа, то h округляют до целого значения; ели x_i содержат десятичные знаки, то h округляют до значения d, содержащего такое же число десятичных знаков. Затем подсчитывается частота n_i, с которой попадают значения x_i признака X в i-й интервал. За начало a_0 первого интервала рекомендуется брать величину

$$a_0 = x_{\min} - \frac{h}{2}.$$

Конец a_k последнего интервала находят по формуле

$$a_k = x_{\max} + \frac{h}{2}.$$

Сформированный интервальный вариационный ряд записывают так:

Интервалы, $(a_{i-1}; a_i)$	$(a_0; a_1)$	$(a_1; a_2)$	...	$(a_{k-1}; a_k)$
Частоты, n_i	n_1	n_2	...	n_k

Интервальный вариационный ряд изображают в виде гистограммы частот n_i или гистограммы *относительных частот* $w_i = \dfrac{n_i}{n}$.

Гистограммой называется ступенчатая фигура, для построения которой по оси абсцисс откладывают отрезки длиной h, изображающие частичные интервалы $(a_{i-1}; a_i)$ варьирования признака X, и на этих отрезках как на основаниях строят прямоугольники с высотами, равными отношению частоты к длине отрезка $\dfrac{n_i}{h}$ или отношению относительной частоты к длине отрезка $\dfrac{n_i}{nh}$ соответствующих интервалов (рис. 1.1). Гистограмма относительных частот является аналогом закона распределения случайной величины X, играет роль *эмпирического закона распределения.*

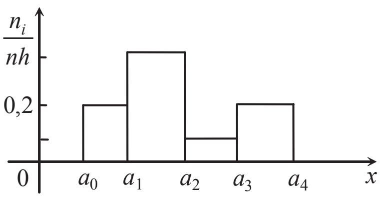


Рис. 1.1. Гистограмма

Для расчета *характеристик* (выборочной средней, выборочной дисперсии, начального и центрального выборочного момента порядка l) переходят от интервального вариационного ряда к дискретному. В качестве вариант x_i этого ряда берут середины интервалов $(a_{i-1}; a_i)$.

Графически дискретный вариационный ряд изображают в *виде полигона частот* (соответственно *полигона относительных частот*) следующим образом. Сначала на числовой плоскости ставят точки $(x_i; n_i)$ (точки $(x_i; w_i)$), где x_i — i-я варианта, n_i — частота (w_i — относительная частота). Затем строят ломаную, соединяющую построенные точки, которую и называют *полигоном* (рис. 1.2).

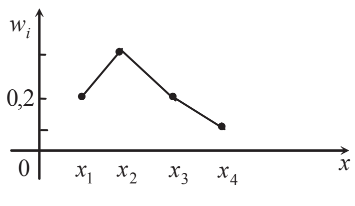


Рис. 1.2. Полигон

Для характеристики свойств статистического распределения в математической статистике вводится понятие эмпирическая функция распределения.

Эмпирической функцией распределения или *функцией распределения* называется функция $F^*(x)$, определяемая равенством

$$F^*(x) = \frac{n_x}{n} = \sum_{i: x_i < x} \frac{n_i}{n},$$

где n — объем выборки; n_x — число вариант x_i, меньших x (рис. 1.3).

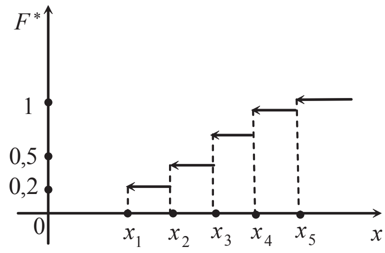


Рис. 1.3. График эмпирической функции распределения

Эмпирическая функция $F^*(x)$ служит для оценки теоретической функции распределения генеральной совокупности. Различие между ними состоит в том, что теоретическая функция $F(x)$ определяет вероятность события $X < x$, а эмпирическая функция $F^*(x)$ определяет относительную частоту этого события.

Теорема Гливенко. *Эмпирическая функция распределения стремится к теоретической по вероятности (сходится по вероятности к $F(x)$) при большом объеме выборки.*

$$\lim_{n\to\infty} P\left(\left|F^*(x) - F(x)\right| < \varepsilon\right) = 1 \quad \forall x, \quad \forall \varepsilon > 0.$$

1.2. Выборочные характеристики статистического распределения

Рассмотрим выборку объемом n со значениями $x_1, x_2, ..., x_n$ признака X. Для характеристики важнейших свойств статистического распределения используют средние показатели, называемые *выборочными числовыми характеристиками*. Если значения x_i признака X не сгруппированы в вариационные ряды и объем выборки n небольшой, то выборочные характеристики находят по следующим формулам:

- $\bar{x} = \frac{1}{n}\sum_i x_i$ — *выборочное среднее* признака X;
- $D_{\text{В}} = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$ — *выборочная дисперсия* признака X.

Если результаты наблюдений сгруппированы в дискретный вариационный ряд, то те же характеристики находят по формулам

$$\overline{x} = \frac{1}{n}\sum_i x_i n_i , \; n = \sum_i n_i ,$$

$$D_{\text{В}} = \frac{1}{n}\sum_i \left(x_i - \overline{x}\right)^2 n_i = \frac{1}{n}\sum_i n_i x_i^{\,2} - \overline{x}^{\,2} . \qquad (*)$$

По формуле (*) вычисляют выборочную дисперсию $D_{\text{В}}$ в случае, если объем выборки $n \geq 30$. Если же $n < 30$, то вычисляют *исправленную выборочную дисперсию* $\hat{D}_{\text{В}}$ по формуле:

- для обыкновенной выборки

$$\hat{D}_{\text{В}} = \frac{1}{n-1}\sum_i \left(x_i - \overline{x}\right)^2 ;$$

- для вариационного ряда

$$\hat{D}_{\text{В}} = \frac{1}{n-1}\sum_i \left(x_i - \overline{x}\right)^2 n_i .$$

Выборочное среднее квадратическое отклонение находят по формулам

$$\sigma_{\text{В}} = \sqrt{D_{\text{В}}} \;\text{ или }\; \hat{\sigma}_{\text{В}} = \sqrt{\hat{D}_{\text{В}}}$$

при различных объемах выборки.

Начальный выборочный момент порядка l вычисляют по формулам:

- для обыкновенной выборки

$$\beta_l^* = \frac{1}{n}\sum_i x_i^{\,l} ;$$

- для вариационного ряда

$$\beta_l^* = \frac{1}{n}\sum_i x_i^{\,l} n_i \;\Rightarrow\; \beta_1^* = \overline{x} .$$

Центральный выборочный момент порядка l определяют по формулам:

- для обыкновенной выборки

$$\mu_l^* = \frac{1}{n}\sum_i \left(x_i - \overline{x}\right)^l ;$$

- для вариационного ряда

$$\mu_l^* = \frac{1}{n}\sum_i \left(x_i - \bar{x}\right)^l n_i \;\Rightarrow\; \mu_2^* = D_{\text{В}}.$$

Для анализа вариационных рядов вычисляют такие характеристики, как моду и медиану. *Модой* Мо называют варианту, которая имеет наибольшую частоту. Например, для вариационного ряда

x_i	37	38	39	40	41	42	43	44	45
n_i	0,8	6,1	10,1	20,1	27,6	22,8	10,1	2,2	0,2

мода $\text{Mo} = 41$. *Медианой* Me называют варианту, которая делит вариационный ряд на равные по числу вариант части. При нечетном объеме выборки $n = 2k+1$ медиана равна $\text{Me} = x_{k+1}$. Например, для вариационного ряда

x_i	2	3	4	5	6
n_i	12	41	54	77	15

медиана $\text{Me} = x_{100} = 4$.

При четном объеме выборки $n = 2k$ медиана находится по формуле

$$\text{Me} = \frac{x_k + x_{k+1}}{2},$$

где x_k — варианта, которая находится слева от середины вариационного ряда, а x_{k+1} — справа от нее. Например, для вариационного ряда

x_i	8	9	10	11	12
n_i	7	10	15	12	6

медиана $\text{Me} = 10$.

1.3. Статистическое распределение. Расчет основных числовых характеристик

Приведем методику проведения лабораторной работы по решению основных задач математической статистики о первичной обработке данных — упорядочение результатов наблюдения или эксперимента, представление их в обозримом виде, определение основных числовых характеристик статистического распределения выборки.

Цель работы:

- научиться вычислять основные числовые характеристики выборки;
- графически изображать вариационный ряд;
- строить эмпирическую функцию распределения выборки.

Задача. По некоторым областям имеются следующие данные о числе преступлений за отчетный период времени.

Число преступлений, тыс.	Менее 20	20–40	40–60	60–80	80–100	Более 100
Число областей	3	6	17	25	10	4

Вычислить выборочное среднее число преступлений по области, моду, медиану, выборочное среднее квадратическое отклонение. Построить гистограмму и график эмпирической функции распределения.

Порядок выполнения работы

Запуск Microsoft Excel

Запустить **Microsoft Excel** и сразу сохранить документ под именем «Задача 1».

Переименуйте *Лист1* в *Расчетная таблица*. Для этого нужно щелкнуть правой мышью по ярлычку этого листа и в открывшемся контекстном меню выбрать команду *Переименовать*. Затем ввести новое имя листа и щелкнуть левой кнопкой мыши вне ярлычка (или нажать клавишу *Enter*). В результате произойдет переименование листа. Переименовать текущий лист можно также с помощью команды *Ячейки/Формат/Переименовать лист* на вкладке *Главная* или двойного щелчка левой кнопкой по ярлычку листа, ввести новое имя. Используя контекстное меню ярлычка рабочего листа, можно также выполнить следующие операции:

- *Вставить* (рабочая книга может быть расширена до 255 листов);
- *Удалить* текущий лист;
- *Переместить* или *Скопировать* текущий рабочий лист (указать, в какую рабочую книгу следует переместить выбранный лист и перед каким листом он должен располагаться).

Удалите остальные листы рабочей книги «Задача 1». Расчетная таблица должна иметь следующий вид:

Интервалы	a_{i-1}	a_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	w_i	f_i	sn_i	sw_i
0–20	0	20		3						
20–40	20	40		6						
40–60	40	60		17						
60–80	60	80		25						
80–00	80	100		10						
100–120	100	120		4						
Итого										

Заполнение расчетной таблицы

Наберите заголовки в первой строке. Чтобы сделать текст нижним или верхним индексом, можно действовать так: сначала набрать в ячейке B1 просто a_{i-1}, затем в строке формул выделить $i - 1$, выбрать на вкладке *Главная Шрифт*, там, в разделе *Видоизменение*, установите галочку напротив надписи «подстрочный». Аналогично делаются индексы в остальных ячейках. Сделайте заголовки полужирным, выравнивание по центру.

Введите данные, известные из условия задачи: частоты n_i, левые границы интервалов a_{i-1}, правые границы a_i.

В столбце D необходимо вычислить середины интервалов x_i по формуле

$$x_i = \frac{a_{i-1} + a_i}{2}.$$

Однако в формулах Excel вместо математических обозначений должны быть адреса ячеек, где находятся соответствующие значения.

Внимание! Не путайте математические формулы с формулами Excel! Помните, что любая формула начинается со знака =, в адресах ячеек используются английские буквы, пробелов делать нельзя.

То есть в ячейку D2 введите следующую формулу:

$$=(B2+C2)/2.$$

В остальные ячейки столбца D копируйте формулы с помощью маркера заполнения.

Далее заполните столбцы F и G по формулам

$$x_i n_i = x_i \cdot n_i,$$

$$x_i^2 n_i = x_i \cdot x_i \cdot n_i$$

соответственно. Для этого в ячейку F2 введите следующую расчетную формулу

$$= D2 * E2,$$

в ячейку G2 — формулу

$$= D2 * D2 * E2 .$$

Остальные ячейки столбцов F и G заполните с помощью маркера заполнения.

В строке *Итого* в столбцах E, F и G вычислите итоговые суммы. Для этого установите курсор мыши на ячейке E8 и нажмите кнопку *Автосумма* на панели инструментов вкладки *Формула*. Ячейки F8 и G8 заполните с помощью маркера заполнения.

Вычисление числовых характеристик статистической выборки

Строку 10 отведем под вычисление ширины интервала. В ячейке A10 наберите текст $h =$, а в B10 введите формулу для вычисления:

$$= C2 - B2 .$$

В следующую строку, в ячейку A11 ввести $x_{\text{ср}} =$. *Выборочное среднее* — среднее арифметическое значений выборки вычисляется по формуле

$$x_{\text{ср}} = \frac{\sum x_i n_i}{\sum n_i} .$$

Суммарные значения произведений $x_i \cdot n_i$ вычислили в ячейке F8, суммарные значения n_i — E8. Таким образом, в ячейку B11 введите следующую формулу:

$$= F8 / E8 .$$

Аналогично в строке 12, в ячейке A12 наберите $x^2{}_{\text{ср}} =$ и вычислите среднее арифметическое квадратов значений выборки $x^2{}_{\text{ср}}$ по формуле

$$x^2{}_{\text{ср}} = \frac{\sum x_i^2 n_i}{\sum n_i} .$$

Следующую строку отведем под *выборочную дисперсию* $D_{\text{В}}$ — среднее значение квадрата отклонения $x_i - x_{\text{ср}}$, которую вычислим по формуле

$$D_{\text{В}} = x^2{}_{\text{ср}} - (x_{\text{ср}})^2 .$$

В формулах Excel для операции возведения в степень используется символ ^ (он получается нажатием одновременно клавиш «Shift» и «6»). Итак, в ячейке A13 введите $D=$, в ячейку B13 введите формулу

$$=B12-B11\hat{}\,2\,.$$

В следующей строке выборочное среднее квадратическое отклонение $\sigma_{\text{В}}$ подпишем как SI (ячейка A14) и вычислим значение по формуле

$$\sigma_{\text{В}} = \sqrt{D_{\text{В}}}\,.$$

Для вычисления корня перейдите в ячейку B14, на панели инструментов вкладки *Формула* нажмите кнопку *Вставить функцию* f_x. В открывшемся окне *Мастера функций* выберите в списке категорий *Математические*, в списке функций, используя полосу прокрутки, выберите *КОРЕНЬ*, щелкните *OK*, далее в окне *Аргументы функции* (рис. 1.4) в поле *Число* введите адрес ячейки B13, нажмите *OK*.

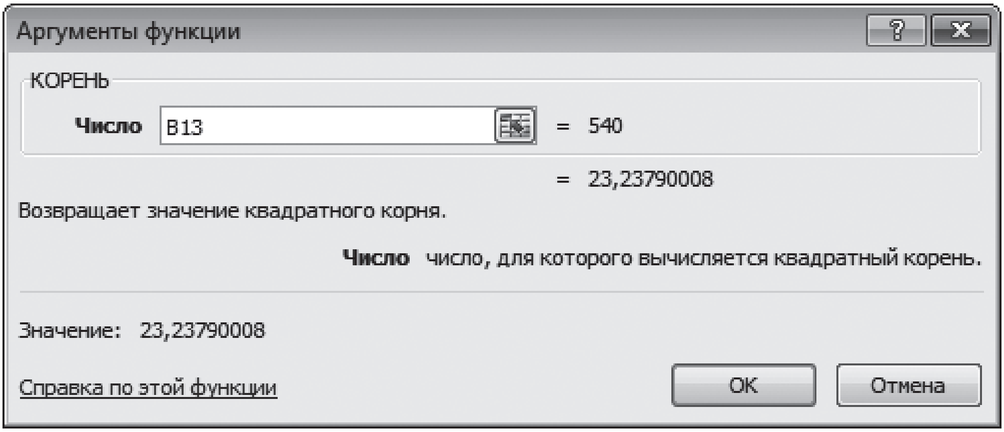


Рис. 1.4. Окно *Аргументы функций*

Заполните столбец *относительных частот* w_i согласно формуле

$$w_i = \frac{n_i}{\sum n_i}\,.$$

В ячейку H2 надо ввести формулу: частоту n_i текущего интервала разделить на объем выборки Σn_i. После чего необходимо копировать эту формулу для остальных интервалов, но при этом не должен меняться адрес клетки, в которой записана общая сумма частот (объем выборки), то есть адрес E8.

Адреса, которые не меняются, называются *абсолютными*. Если необходимо сделать адрес ячейки абсолютным, то поставьте перед ним знак доллара \$. Таким образом, в ячейку H2 введите формулу

$$= E2 / \$E\$8 .$$

Для проверки подсчитайте итоговую сумму в столбце H, которая должна быть равна 1.

В следующем столбце J *плотности относительной частоты* f_i вычисляются по формуле

$$f_i = \frac{w_i}{h} .$$

В ячейку J2 введите формулу

$$= H2 / \$B\$10 .$$

Для удобства вычисления накопленных частот sn_i и накопленных относительных частот sw_i, а также для построения графика эмпирической функции распределения добавьте строку в начало таблицы (между 1-й и 2-й строками). Для этого выделите *строку* 2 и выберите в контекстном меню команду *Вставить*.

В ячейки J2 и K2 введите 0, так как накопление начинается с 0. В ячейку J3 введите формулу

$$= J2 + E3 ,$$

в ячейку K3 — формулу

$$= K2 + H3 .$$

В остальные ячейки столбцов J и K копируйте формулы с помощью маркера заполнения.

Мода, медиана

Строки 16 и 17 отведем соответственно под *моду* и *медиану*. В ячейки A16 и A17 введите Mo= и Me=.

Модальным является интервал с наибольшей частотой.

Моду вычисляем по формуле

$$\mathrm{Mo} = x_{\mathrm{Mo}} + \frac{n_{\mathrm{Mo}} - n_{\mathrm{Mo}-1}}{n_{\mathrm{Mo}} - n_{\mathrm{Mo}-1} + n_{\mathrm{Mo}} - n_{\mathrm{Mo}+1}} h,$$

где x_{Mo} — начало модального интервала; n_{Mo} — частота модального интервала; $n_{\mathrm{Mo}-1}$ — частота интервала, предшествующего модальному; $n_{\mathrm{Mo}+1}$ — частота интервала, следующего за модальным; h — длина интервала.

По расчетной таблице определяем, что модальный интервал — это интервал 60–80 с частотой, равной 25.

Таким образом, в ячейку B16 необходимо ввести формулу

$$= D6 + (E6 - E5) / (E6 - E5 + E6 - E7) * B11.$$

Медианным называется интервал, для которого накопленная частота sn_i впервые превысит половину суммы частот $\sum n_i / 2$.

Медиану вычислим по формуле

$$\mathrm{Me} = x_{\mathrm{Me}} + \frac{\sum n_i / 2 - sn_{\mathrm{Me}-1}}{n_{\mathrm{Me}}} h,$$

где x_{Me} — начало медианного интервала; n_{Me} — частота медианного интервала; $sn_{\mathrm{Me}-1}$ — накопленная частота интервала, предшествующего медианному; h — длина интервала.

По расчетной таблице определяем величину $\sum n_i / 2$: ее значение равно 32,5, поэтому медианным является интервал 60–80 с частотой, равной 25. То есть в ячейку B17 необходимо ввести формулу

$$= D6 + (E9 / 2 - J5) / E6 * B11.$$

Построение диаграмм

Программа MS Excel дает возможность представлять числовую информацию в графической форме — в виде диаграмм и графиков. Для создания диаграмм в табличный процессор встроены специальные средства, позволяющие просто и наглядно выполнить необходимые операции: выбрать тип диаграммы, задать диапазон ячеек, по значениям которых строится диаграмма, ввести название диаграммы, обозначения осей, указать масштабы, условные обозначения элементов, выбрать цвета надписей, заливки элементов и т. д.

Гистограмма

Построим гистограмму относительных частот вариационного ряда. *Гистограммой относительных частот* называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высоты равны плотностям относительной частоты f_i. Площадь i-го частичного прямоугольника

$$hf_i = h(w_i / h) = w_i\,,$$

где w_i — относительная частота вариант, попавших в i-й интервал. Площадь гистограммы относительных частот равна сумме всех относительных частот, то есть единице.

На вкладке *Вставка* в строке меню *Диаграммы* выбрать вид диаграммы *Гистограмма,* затем подвид диаграммы *Гистограмма с группировкой* (слева в первом ряду на рис. 1.5).

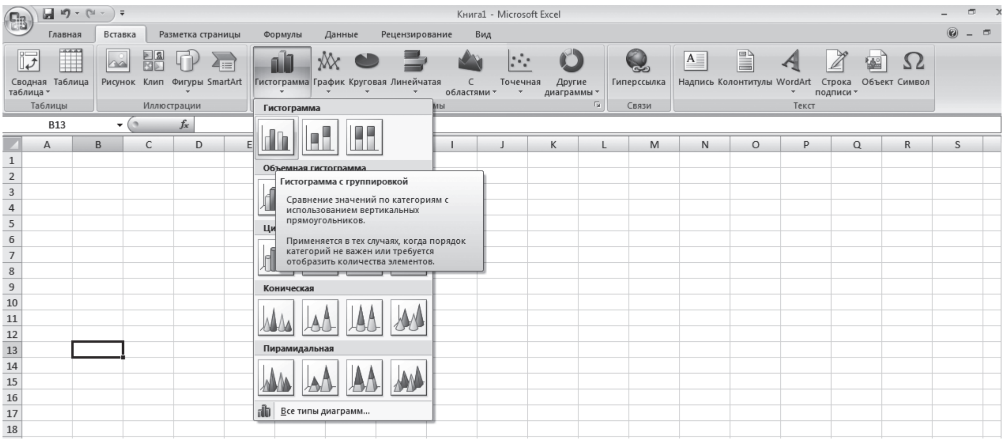


Рис. 1.5. Вид диаграммы

Выбрать тип диаграммы можно также другим способом: на вкладке *Вставка* нажать на кнопку вызова  справа от строки меню *Диаграммы*, после чего в открывшемся окне *Вставка диаграммы* из полного списка доступных диаграмм и графиков выбрать требуемый вариант диаграммы (рис. 1.6).

В результате на рабочем листе появится окно с пустой диаграммой. Для отображения на ней данных выполнить следующие действия:

1. В контекстном меню диаграммы выполнить пункт *Выбрать данные.*

2. В открывшемся окне в поле *Диапазон данных для диаграммы* указать ячейки, в которых находятся плотности относительных частот f_i. В разделе *Подписи горизонтальной оси* нажать *Изменить* и указать в поле *Диапазон подписей оси* ячейки, в которых находятся средины интервалов x_i . Нажмите *OK*.

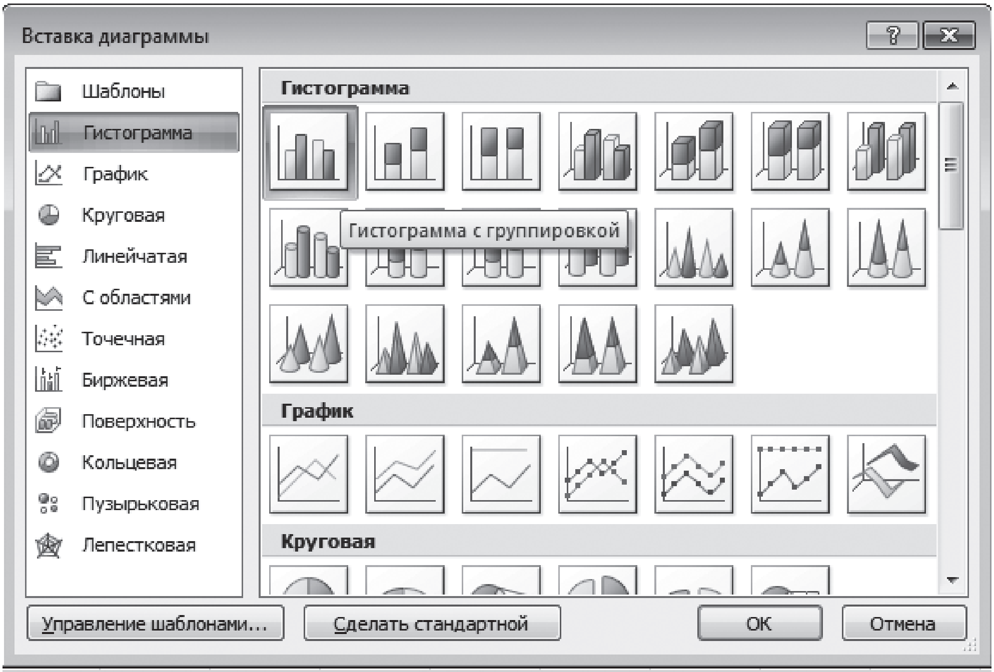


Рис. 1.6. Вставка диаграммы

Преобразование диаграмм

Для внесения изменений в диаграмму используется командная строка *Работа с диаграммами* (рис. 1.7), а также контекстные меню элементов самой диаграммы. Щелкнуть по диаграмме, на вкладке *Макет* в группе *Фон* нажать на кнопку *Область построения*, выбрать в выпадающем меню вариант *Нет* (*удалить заливку области построения*).

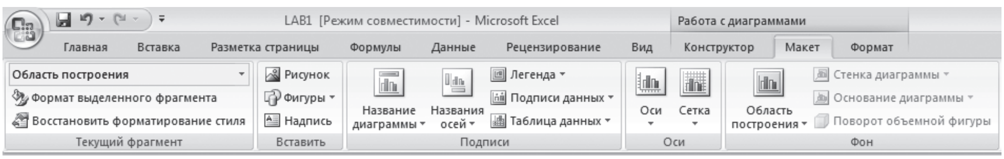


Рис. 1.7. Командная строка *Работа с диаграммами*

Нажать правую клавишу мыши на одном из столбцов построенной гистограммы, выбрать команду *Формат ряда данных*, пункт *Параметры ряда*, там установить *Боковой зазор* 0 %, в пункте *Заливка* установить флажок *Разноцветные точки*.

Щелкнуть по кнопке *Названия осей*, далее выбрать *Название основной горизонтальной оси*, затем вариант *Название под осью*. Вместо *Названия оси* ввести *Интервалы*.

Добавить заголовок диаграммы «Гистограмма». Для этого нажать кнопку *Название диаграммы* группы команд *Подписи* вкладки *Макет*, выбрать в появившемся меню вариант *Над диаграммой*, изменить название на *Гистограмма*. Затем установить шрифт для заголовка: размер 18, начертание полужирный курсив, цвет синий.

Поместить диаграмму на отдельном листе с именем *Гистограмма*. Для этого щелкнуть диаграмму, на вкладке *Конструктор* в группе *Расположение* нажать кнопку *Переместить диаграмму*. В появившемся окне (рис. 1.8) в разделе *Разместить диаграмму* установить переключатель в положение *На отдельном листе* и ввести имя нового листа диаграммы в поле этого переключателя. Нажать *OK*.

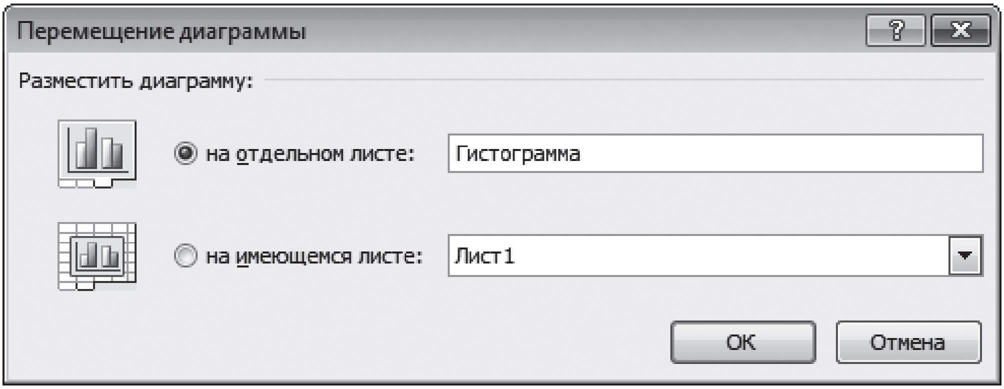


Рис. 1.8. Перемещение диаграммы

На рис. 1.9 приведена гистограмма, которая должна получиться.

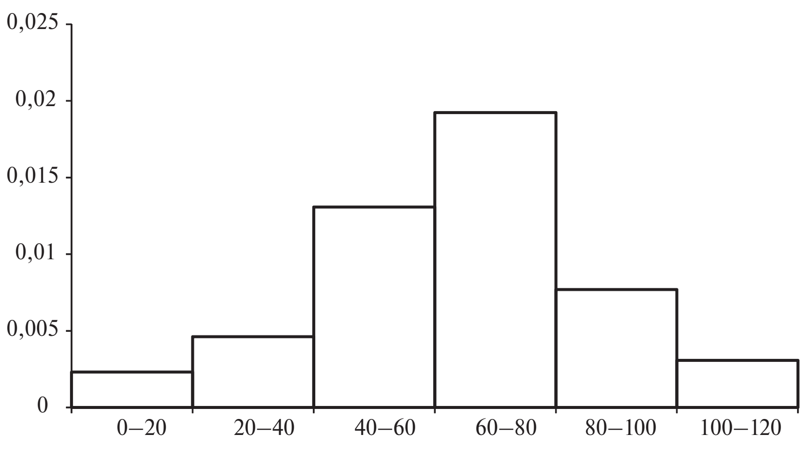


Рис. 1.9. Гистограмма

Эмпирическая функция распределения

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$:

$$F^*(x) = \frac{n_x}{\sum n},$$

где n_x — число вариант меньших *x*.

Для построения эмпирической функции распределения задачи добавить в начало расчетной таблицы (перед первым столбцом) столбец с названием *узлы*. Набрать в ячейке A2 значение начала первого интервала 0, в ячейке A3 набрать 20, далее выделить ячейки A2 и A3, подвести указатель мыши к маркеру заполнения и, удерживая левую кнопку мыши, потащить вниз до конца столбца таблицы.

На вкладке *Вставка* в строке меню *Диаграммы* выбрать вид диаграммы *Точечная,* затем подвид диаграммы *Точечная с прямыми отрезками и маркерами* (слева во втором ряду на рис. 1.10).

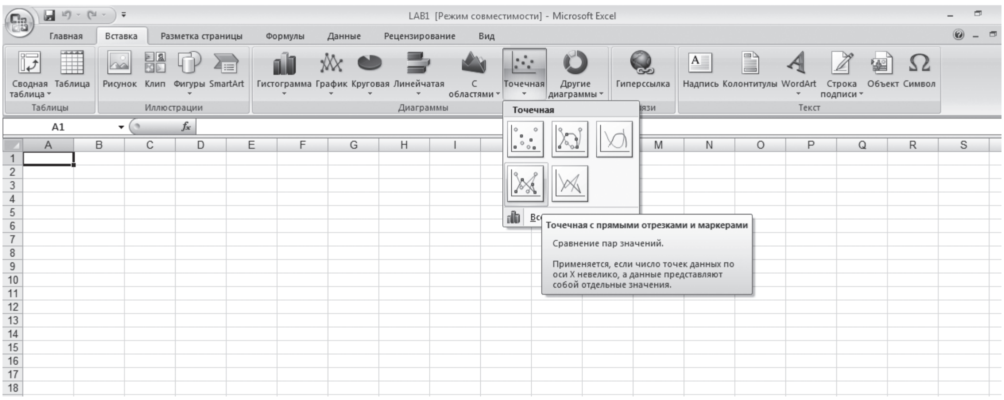


Рис. 1.10. Точечная диаграмма

В результате на рабочем листе появится окно с пустой диаграммой. Для отображения на ней данных выполните следующие действия:

1. В контекстном меню диаграммы выполнить пункт *Выбрать данные*.

2. В открывшемся окне в поле *Диапазон данных для диаграммы* указать ячейки, в которых находятся накопленные относительные частоты sw_i. В разделе *Элементы легенды* нажать *Изменить* и указать в поле *Значения X* ячейки, в которых находятся узлы. Нажмите *OK*.

Редактирование диаграммы

Расположить диаграмму на отдельном листе с именем *Эмпирическая*. Для этого щелкнуть диаграмму, вызвать контекстное меню и выполнить пункт *Переместить диаграмму*. В появившемся окне (см. рис. 1.8) в разделе *Разместить диаграмму* установить переключатель в положение *На отдельном листе* и ввести имя нового листа диаграммы в поле этого переключателя. Нажать *OK*.

Щелкнуть по диаграмме, в командной строке *Работа с диаграммами* (см. рис. 7) на вкладке *Макет* в группе *Фон* нажать на кнопку *Область построения*, выбрать в выпадающем меню вариант *Нет* (*удалить заливку области построения*).

В группе *Оси* кнопки *Оси* и *Сетка* позволяют настроить отображение осей и линий сетки для диаграммы. Опции *Основная горизонтальная ось* и *Основная вертикальная ось* выбираются по умолчанию. Создать и горизонтальные, и вертикальные линии сетки для основных делений — опция *Основные линии сетки* (рис. 1.11).

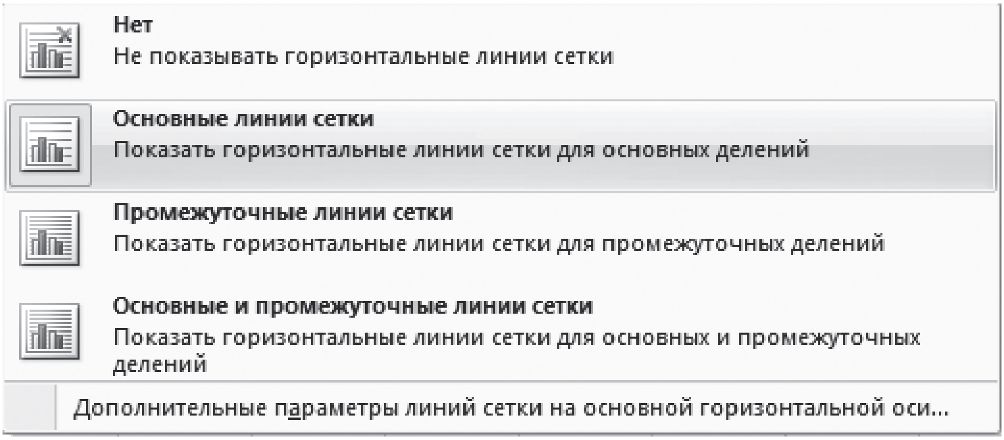


Рис. 1.11. Параметры линий сетки диаграммы

Щелкнуть кнопку *Названия осей*, далее выбрать *Название основной горизонтальной оси*, затем вариант *Название под осью*. Ввести вместо *Названия оси* обозначение x, затем перетащить его (методом «буксировки») в положение справа от оси. Выполнить аналогичные действия для названия основной вертикальной оси, введя обозначение для нее $F^*(x)$ и перетащив его в положение над осью. Щелкнув кнопку *Легенда*, выбрать вариант *Нет*.

Добавить заголовок диаграммы «Эмпирическая функция распределения». Для этого нажать кнопку *Название диаграммы* группы команд *Подписи* вкладки *Макет*, выбрать в появившемся меню вариант *Над диаграммой*, изменить название на *Эмпирическая функция распределения*. Затем установить шрифт для заголовка: размер 18, начертание полужирный, цвет синий.

Вызвать контекстное меню (щелкните правой кнопкой мыши) построенной линии, выбрать команду *Формат ряда данных*, пункт *Цвет линии*, установить *красный*, в пункте *Заливка маркера* установить цвет *красный* (рис. 1.12).

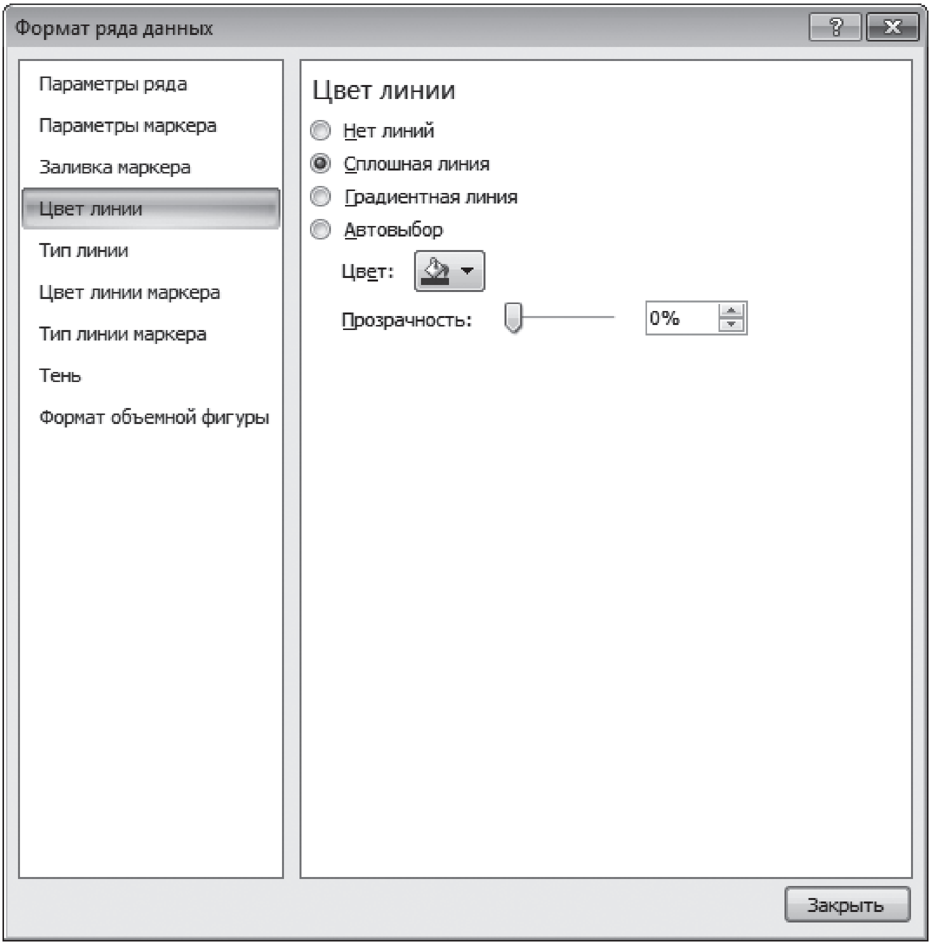


Рис. 1.12. Формат ряда данных

Добавить значения данных на диаграмме. Для этого вызвать контекстное меню на диаграмме, выберите пункт *Добавить подписи данных*.

Для настройки параметров горизонтальной и вертикальной осей (форматирование этих элементов диаграммы) следует перейти к окну *Формат оси* с помощью контекстного меню (рис. 1.13). Для горизонтальной оси выбрать на вкладке *Параметры оси* максимальное значение 120, цену основных делений 20. Для вертикальной оси выбрать минимальное значение 0, максимальное значение 1, цену основных делений 0,2; на вкладке *Число* установить формат *Числовой*, число десятичных знаков 2.

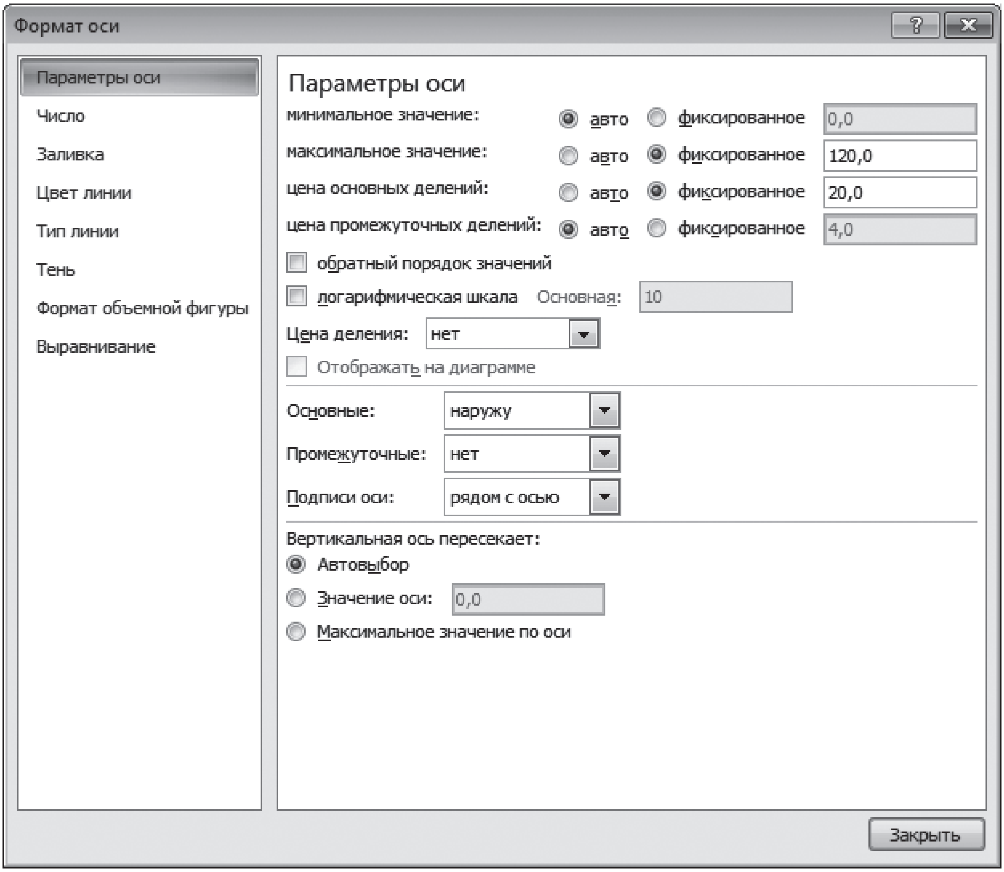


Рис. 1.13. Формат оси

В результате должен получиться график эмпирической функции распределения, как на рис. 1.14.

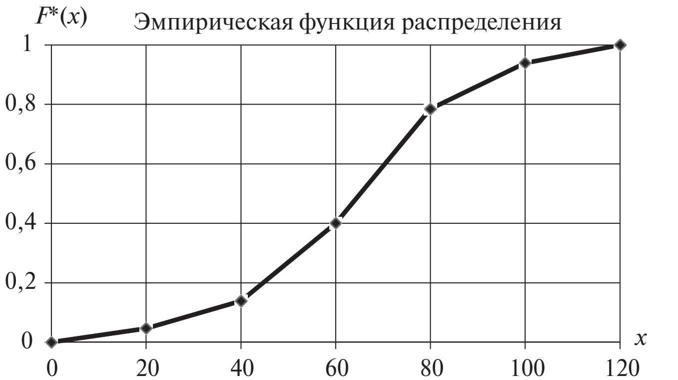


Рис. 1.14. График эмпирической функции распределения

Индивидуальное задание

Создать лист *Расчетная таблица №*. Решить свой вариант задания.

Результаты работы

В результате выполненной работы студент должен продемонстрировать преподавателю готовый файл *Задача 1.xlsx*, содержащий 7 листов:

- лист 1 (*Титульный лист*) — титульный лист к работе, на котором указаны название работы, номер варианта, Ф. И. О. студента, номер группы, дата выполнения работы;
- лист 2 (*Расчетная таблица*) — таблица исходных данных и решения задачи;
- лист 3 (*Гистограмма*) — гистограмма задачи;
- лист 4 (*Эмпирическая*) — график эмпирической функции распределения задачи;
- лист 5 (*Расчетная таблица №...*) — таблица исходных данных и решения задания своего варианта;
- лист 6 (*Гистограмма №...*) — гистограмма задания своего варианта;
- лист 7 (*Эмпирическая №...*) — график эмпирической функции распределения задания своего варианта.

Варианты индивидуального задания

Вариант 1

На ткацкой фабрике из 1000 ткачих произведена собственно-случайная бесповторная выборка 100 человек. В результате получены следующие данные о распределении ткачих по уровню дневной выработки:

Уровень дневной выработки, м	30–40	40–50	50–60	60–70
Число ткачих, чел.	30	33	24	13

Найти средний уровень дневной выработки, среднее квадратическое отклонение, моду и медиану. Построить гистограмму относительных частот и эмпирическую функцию распределения.

Вариант 2

Получены следующие данные выборки о распределении рабочих предприятия по заработной плате:

Зарплата, тыс. руб.	20–30	30–40	40–50	50–60	60–70	70–80
Число рабочих, чел.	8	19	28	32	42	21

Найти среднюю зарплату по предприятию, среднее квадратическое отклонение, моду и медиану. Построить гистограмму относительных частот и эмпирическую функцию распределения.

Вариант 3

Имеются следующие данные выборки о распределении рабочих завода по стажу работы:

Стаж работы, лет	0–5	5–10	10–15	15–20	20–25	25–30
Число рабочих, чел.	20	30	40	50	40	20

Найти средний стаж работы по заводу, среднее квадратическое отклонение, моду и медиану. Построить гистограмму относительных частот и эмпирическую функцию распределения.

Вариант 4

В результате исследования, проведенного с целью обследования жилищных условий жителей города, получены следующие данные:

Общая площадь на 1 чел., м^2	До 5	5–10	10–15	15–20	20–25	25–30	30 и более
Число жителей, чел.	8	95	204	270	210	130	83

Найти среднюю площадь на 1 человека, среднее квадратическое отклонение, моду и медиану. Построить гистограмму относительных частот и эмпирическую функцию распределения.

Вариант 5

В результате выборочного наблюдения в регионе получены следующие данные о распределении осужденных по срокам свободы:

Сроки лишения свободы, лет	До 2	2–4	4–6	6–8	8–10	10–12	Более 12
Число осужденных, чел.	13	15	14	12	9	5	4

Найти средний срок лишения свободы, среднее квадратическое отклонение, моду и медиану. Построить гистограмму относительных частот и эмпирическую функцию распределения осужденных по срокам лишения свободы.

Вариант 6

Распределение грузов, перевозимых автотранспортным предприятием, характеризуется следующими данными:

Расстояние перевозок, км	До 50	50–100	100–150	150–200	200–250	250–300	Более 300
Количество грузов, % к итогу	23,5	21,1	17,1	13,8	11,6	6,1	6,8

Построить гистограмму относительных частот и эмпирическую функцию распределения. Вычислить среднее расстояние перевозок, медиану, моду, дисперсию.

Вариант 7

Получены следующие данные о распределении продавцов магазина по выработке:

Выработка продавцов, тыс. руб.	80–100	100–120	120–140	140–160	160–180
Число продавцов, чел.	5	10	20	10	5

Построить гистограмму относительных частот и эмпирическую функцию распределения. Вычислить среднюю выработку по магазину, медиану, моду, среднее квадратическое отклонение.

Вариант 8

Имеются результаты измерения роста случайно отобранных 100 студентов.

Рост, см	158–162	162–166	166–170	170–174	174–178	178–182	182–186
Число студентов, чел.	10	14	26	28	12	8	2

Построить гистограмму относительных частот и эмпирическую функцию распределения роста студентов. Найти средний рост, медиану, моду, среднее квадратическое отклонение.

Вариант 9

По данному распределению выборки построить гистограмму относительных частот и эмпирическую функцию распределения. Найти среднее значение выборки, медиану, моду, среднее квадратическое отклонение.

Частичный интервал	1–5	5–9	9–13	13–17	17–21
Частота вариант интервала	10	20	50	12	8

Вариант 10

По данному распределению выборки построить гистограмму относительных частот и эмпирическую функцию распределения. Найти среднее значение выборки, медиану, моду, среднее квадратическое отклонение.

Частичный интервал	10–15	15–20	20–25	25–30	30–35
Частота вариант интервала	2	4	8	4	2

2. Выборочное наблюдение

При статистическом исследовании явлений производится наблюдение не всех единиц изучаемой генеральной совокупности, а лишь некоторой ее части — выборки, и по этой части судят о совокупности в целом. Отбор единиц из генеральной совокупности должен производиться таким образом, чтобы выборочная совокупность была *репрезентативной* (*представительной*), то есть достаточно хорошо характеризовала генеральную совокупность. Выборочные характеристики распространяются затем на всю генеральную совокупность.

Для того чтобы выборка была репрезентативной, должны выполняться следующие условия:

- случайный отбор элементов совокупности;
- равновероятность попадания в выборку любого элемента генеральной совокупности;
- достаточно большой объем выборки.

Существует несколько способов отбора, обеспечивающих репрезентативность выборки. Рассмотрим следующие из них:

- собственно-случайный;
- механический;
- типический;
- серийный.

Собственно-случайный способ заключается в отборе единиц из генеральной совокупности наугад, при этом все единицы имеют равную вероятность попасть в выборку. Собственно-случайный отбор может быть как *повторным* (выбранная для обследования единица возвращается обратно в генеральную совокупность и может участвовать в следующем отборе), так и *бесповторным* (выбранная единица в генеральную совокупность не возвращается и в дальнейшем в отборе не участвует).

Механический способ заключается в отборе единиц из генеральной совокупности с заданной цикличностью. Так, если осуществляется 2%-ный отбор, то в выборку попадает каждая 50-я единица генеральной совокупности ($1/50 \cdot 100\,\% = 2\,\%$); если 10%-ный отбор, то каждая 10-я единица ($1/10 \cdot 100\,\% = 10\,\%$) и так далее. Также механический способ можно организовать, если генеральную совокупность разделить на столько групп, сколько единиц должно войти в выборку, и из каждой группы отобрать случайным образом единицу.

Типический способ отбора применяется в случаях, когда генеральную совокупность можно разбить на однородные типические группы. Из каждой группы производится пропорционально ее объему выборка единиц совокупности собственно-случайным способом (с повторным или бесповторным отбором).

Серийный способ отбора применяется в случаях, когда единицы генеральной совокупности объединены в отдельные группы (серии), имеющие, как правило, равные объемы. Например, в качестве серий можно рассматривать упаковки с определенным количеством готовой продукции, партии однотипного товара, бригады и тому подобное. Этот способ заключается в собственно-случайном (повторном или бесповторном) отборе серий из генеральной совокупности, внутри которых производится сплошное обследование единиц.

2.1. Точечные оценки и их свойства

Основная задача математической статистики состоит в нахождении закона распределения случайной величины X на основании экспериментальных данных. Во многих случаях вид закона распределения X можно считать известным, неизвестны лишь значения параметров распределения ϑ, и задача сводится к получению приближенных значений (оценок) неизвестных параметров ϑ этого распределения.

Точечной оценкой параметра ϑ по выборке $\{x_1, x_2, ..., x_n\}$ называется любая функция $\vartheta^* = \vartheta^*(x_1, x_2, ..., x_n)$, используемая в приближенном равенстве:

$$\vartheta \approx \vartheta^*.$$

Качество оценки определяется следующими свойствами:

- несмещенность;
- состоятельность;
- эффективность.

Оценка ϑ^* называется *несмещенной оценкой* параметра ϑ, если ее математическое ожидание при любом фиксированном n равно оцениваемому параметру ϑ:

$$M\left(\vartheta^*\right) = \vartheta.$$

Несмещенность означает отсутствие систематической ошибки при выборочном наблюдении.

Оценка ϑ^* называется *состоятельной оценкой* параметра ϑ, если она сходится по вероятности к оцениваемому параметру ϑ, то есть

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} P\left(\left|\vartheta^* - \vartheta\right| < \varepsilon\right) = 1 .$$

Из неравенства Чебышева следует, что несмещенная оценка, дисперсия которой стремится к нулю при $n \to \infty$, является состоятельной.

Эффективной называется оценка ϑ^*, которая при фиксированном объеме выборки n является несмещенной и имеет наименьшую возможную дисперсию среди всех несмещенных оценок ϑ^{**} того же самого параметра ϑ:

$$D(\vartheta^*) = \min_{\vartheta^{**}} D(\vartheta^{**}) .$$

Эффективная оценка обладает наименьшим возможным среднеквадратическим отклонением $\sigma(\vartheta^*)$ относительно неизвестного ϑ:

$$\sigma(\vartheta^*) = \sqrt{M\left(\vartheta^* - M(\vartheta^*)\right)^2} = \sqrt{M\left(\vartheta^* - \vartheta\right)^2} \to \min_{\vartheta^{**}} \sigma(\vartheta^{**}) .$$

На рис. 2.1 представлены различные (по качеству) оценки одного и того же неизвестного параметра ϑ.

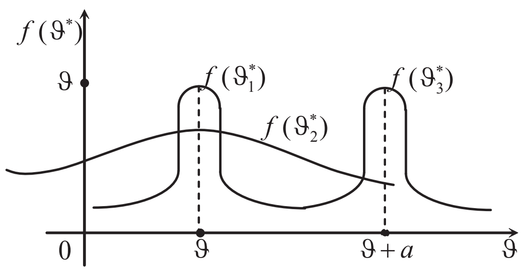


Рис. 2.1. Оценки параметра ϑ:
ϑ_1^*, ϑ_2^* — несмещенные оценки параметра ϑ (ϑ_1^* — более эффективная, чем ϑ_2^*);
ϑ_3^* — смещенная оценка

2.2. Точечные оценки математического ожидания и дисперсии

Получим точечные оценки основных числовых характеристик случайной величины по выборке.

Задача 1

Исследуется случайная величина *X*. Требуется найти оценку ее математического ожидания $M(X)$.

Решение

Будем считать, что математическое ожидание $M(X)=M$ случайной величины *X* является искомым неизвестным параметром ϑ. Найдем оценку ϑ по выборке $\{x_1, x_2, ..., x_n\}$.

Пусть $\vartheta_1^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$ — выборочное среднее.

Оценка ϑ_1^* является функцией выборки $\vartheta_1^* = \vartheta_1^*(x_1, x_2, ..., x_n)$.

Выясним, насколько качественной является замена ϑ на ϑ_1^*:

- *несмещенность.*

Вычислим математическое ожидание оценки ϑ_1^*:

$$M(\vartheta_1^*) = M\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} M(x_i) = \frac{1}{n}\cdot n\cdot M = M = \vartheta .$$

Следовательно, по определению $\vartheta_1^* = \overline{x}$ — несмещенная оценка.

- *состоятельность.*

Рассмотрим $\forall\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\vartheta_1^* - \vartheta\right| < \varepsilon\right) = \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - M\right| < \varepsilon\right),$$

по теореме Чебышева следует, что $\lim\limits_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - M\right| < \varepsilon\right) = 1$,

окончательно имеем $\forall\varepsilon > 0$ $\lim\limits_{n\to\infty} P\left(\left|\vartheta_1^* - \vartheta\right| < \varepsilon\right) = 1$.

Следовательно, по определению $\vartheta_1^* = \overline{x}$ — состоятельная оценка.

Задача 2

Исследуется случайная величина *X*. Требуется найти оценку ее дисперсии $D(X)$.

Решение

Будем считать, что дисперсия $D(X)=D$ случайной величины *X* является искомым неизвестным параметром ϑ. Найдем оценку ϑ по выборке $\{x_1, x_2, ..., x_n\}$.

Пусть $\vartheta_2^* = D_\text{В} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — выборочная дисперсия. Так как математическое ожидание оценки ϑ_2^*

$$M\left(\vartheta_2^*\right) \neq \vartheta,$$

то $D_\text{В}$ — *смещенная оценка* дисперсии случайной величины.

Пусть $\vartheta_3^* = \widehat{D}_\text{В} = \widehat{\sigma}_\text{В}{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — исправленная выборочная дисперсия. Математическое ожидание оценки ϑ_3^*:

$$M\left(\vartheta_3^*\right) = \widehat{D}_\text{В} = \vartheta.$$

Следовательно, по определению $\vartheta_3^* = \widehat{D}_\text{В} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — *несмещенная оценка* дисперсии случайной величины.

Таким образом, выборочные характеристики используются как статистические оценки соответствующих характеристик генеральной совокупности:

- выборочное среднее $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ — несмещенная состоятельная оценка математического ожидания генеральной совокупности;
- выборочная дисперсия $D_\text{В} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — смещенная оценка дисперсии генеральной совокупности;
- исправленная выборочная дисперсия $\widehat{D}_\text{В} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — несмещенная оценка дисперсии генеральной совокупности.

2.3. Методы получения точечных оценок параметров распределения

Приведем основные методы получения точечных оценок.

Метод моментов

Пусть функция распределения случайной величины X зависит от s теоретических параметров: $\vartheta_1, \vartheta_2, ..., \vartheta_S$.

1. Составим систему:

$$\begin{cases} \beta_1 = \beta_1(\vartheta_1, \vartheta_2, ..., \vartheta_s), \\ \beta_2 = \beta_2(\vartheta_1, \vartheta_2, ..., \vartheta_s), \\ \dots\dots\dots\dots\dots\dots \\ \beta_S = \beta_s(\vartheta_1, \vartheta_2, ..., \vartheta_s), \end{cases}$$

где β_l — начальный момент случайной величины X порядка l:

$$\beta_l = M\left(X^l\right).$$

2. Предположим, что возможно обращение системы, то есть изменение ролей между функциями и аргументами:

$$\begin{cases} \vartheta_1 = \vartheta_1(\beta_1, \beta_2, ..., \beta_s), \\ \vartheta_2 = \vartheta_2(\beta_1, \beta_2, ..., \beta_s), \\ \dots\dots\dots\dots\dots\dots \\ \vartheta_s = \vartheta_s(\beta_1, \beta_2, ..., \beta_s). \end{cases}$$

3. По выборке $\{x_1, x_2, ..., x_n\}$ находим начальные выборочные моменты:

$$\beta_l^* = \frac{1}{n}\sum_{i=1}^{n} x_i^l .$$

4. Подставляем полученные в пункте 3 значения в систему 2, решая систему относительно неизвестных $\vartheta_1, \vartheta_2, ..., \vartheta_S$, находим точечные оценки неизвестных параметров распределения $\vartheta_1^*, \vartheta_2^*, ..., \vartheta_s^*$:

$$\vartheta_i^* = \vartheta_i\left(\beta_1^*, \beta_2^*, ..., \beta_s^*\right), \ i = 1, 2, ..., s .$$

Задача 3

Пусть X — случайная величина с нормальным законом распределения, но параметры a и σ^2 — неизвестны. Оценить их по методу моментов.

Решение

1. Составляем систему, связывающую начальные моменты 1-го и 2-го порядка с неизвестными параметрами распределения случайной величины X:

$$\begin{cases} \beta_1 = M(X) = a, \\ \beta_2 = M(X^2) = \sigma^2 + a^2. \end{cases}$$

2. Обращаем систему:

$$\begin{cases} a = \beta_1, \\ \sigma^2 = \beta_2 - \beta_1^2. \end{cases}$$

3. Находим начальные выборочные моменты:

$$\begin{cases} \beta_1^* = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}, \\ \beta_2^* = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2. \end{cases}$$

4. Подставляем полученные значения выборочных моментов в систему пункта 2:

$$\begin{cases} a^* = \beta_1^*, \\ \left(\sigma^2\right)^* = \beta_2^* - \left(\beta_1^*\right)^2. \end{cases}$$

Путем несложных выкладок получаем:

$$\left(\sigma^2\right)^* = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = D_{\text{В}}.$$

Итак, получены искомые точечные оценки параметров нормального закона распределения:

$$a^* = \overline{x},$$

$$\left(\sigma^2\right)^* = D_{\text{В}}.$$

Метод максимального правдоподобия

Пусть $\vartheta_1, \vartheta_2, ..., \vartheta_S$ — неизвестные параметры распределения.

Точечные оценки $\vartheta_1^*, \vartheta_2^*, ..., \vartheta_s^*$, доставляющие максимум функции правдоподобия $L(\vartheta_1, \vartheta_2, ..., \vartheta_s)$, называются *точечными оценками метода максимального правдоподобия*.

Функция $L = L(\vartheta_1, \vartheta_2, ..., \vartheta_s)$ является законом распределения выборки $\{X_1, X_2, ..., X_n\}$ при определенных значениях результатов измерений $\{x_1, x_2, ..., x_n\}$.

Для *дискретных случайных величин*:

$$L(\vartheta_1, \vartheta_2, ..., \vartheta_s) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) =$$
$$= P(X_1 = x_1) \cdot P(X_2 = x_2) \cdot ... \cdot P(X_n = x_n). \qquad (2.1)$$

Здесь в каждом множителе содержатся искомые параметры $\vartheta_1, \vartheta_2, ..., \vartheta_S$.

Для *непрерывных случайных величин*:

$$L(\vartheta_1, \vartheta_2, ..., \vartheta_s) = f(x_1; \vartheta_1, \vartheta_2, ..., \vartheta_s) \cdot f(x_2; \vartheta_1, \vartheta_2, ..., \vartheta_S) \cdot ... \cdot f(x_n; \vartheta_1, \vartheta_2, ..., \vartheta_S). \qquad (2.2)$$

Искомые оценки ϑ_i^* удобно находить из необходимого условия экстремума функции правдоподобия $L(\vartheta_1, \vartheta_2, ..., \vartheta_s)$, но так как функция $L(\vartheta_1, \vartheta_2, ..., \vartheta_s)$ достигает максимума в тех же точках, что и логарифмическая функция правдоподобия $\ln L(\vartheta_1, \vartheta_2, ..., \vartheta_s)$, то оценки ϑ_i^* находим, решая систему уравнений

$$\frac{\partial}{\partial \vartheta_i} \ln L(\vartheta_1, \vartheta_2, ..., \vartheta_s) = 0, \; i = 1, 2, ..., s.$$

Задача 4

Случайная величина X — распределена по нормальному закону. Оценить параметры a и σ^2 по методу максимального правдоподобия.

Решение

Пусть $\{x_1, x_2, ..., x_n\}$ — выборка наблюдений случайной величины X с плотностью распределения

$$f(x, a, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Следовательно, неизвестны параметры $\begin{cases} \vartheta_1 = a \\ \vartheta_2 = \sigma^2. \end{cases}$

Найдем функцию правдоподобия $L(a,\sigma^2)$. По формуле (2.2) имеем:

$$L(a,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{(x_1-a)^2}{2\sigma^2}} e^{-\frac{(x_2-a)^2}{2\sigma^2}} \cdot \ldots \cdot e^{-\frac{(x_n-a)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^{n}(x_i-a)^2}{2\sigma^2}}.$$

Логарифмическая функция правдоподобия отсюда равна:

$$\ln L(a,\sigma^2) = -\frac{n}{2}\left(\ln 2\pi + \ln\sigma^2\right) - \frac{\sum_{i=1}^{n}(x_i-a)^2}{2\sigma^2}.$$

Используя необходимые условия экстремума $\ln L(a,\sigma^2)$, получим систему уравнений для нахождения оценок:

$$\begin{cases} \dfrac{\partial \ln L}{\partial a} = \dfrac{2\sum_{i=1}^{n}(x_i-a)}{2\sigma^2} = 0, \\ \dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{\sum_{i=1}^{n}(x_i-a)^2}{2\sigma^4} = 0 \end{cases}.$$

Из первого уравнения системы следует: $\sum_{i=1}^{n} x_i = na \Rightarrow a^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$.

Из второго уравнения и из оценки a^* имеем

$$\frac{\sum_{i=1}^{n}(x_i-a)^2}{2\sigma^4} = \frac{n}{2\sigma^2} \Rightarrow \sum_{i=1}^{n}(x_i-a)^2 = n\sigma^2 \Rightarrow (\sigma^2)^* = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = D_{\text{В}}.$$

Задача 5

Случайная величина X имеет закон распределения Пуассона с параметром λ. Оценить по методу максимального правдоподобия параметр λ.

Решение

Пусть $\{x_1, x_2, ..., x_n\}$ — выборка наблюдений случайной величины X, имеющей распределение Пуассона с неизвестным параметром λ:

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda},$$

где x принимает неотрицательные целочисленные значения $x = 0, 1, 2, \ldots$.

Следовательно, неизвестный параметр $\vartheta_1 = \lambda$.

Функция правдоподобия $L(\lambda)$ выборки объема n определяется по формуле (2.1):

$$L(\lambda) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!} e^{-\lambda} \cdot \ldots \cdot \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1!\, x_2! \ldots x_n!} e^{-\lambda n}.$$

Найдем логарифмическую функцию правдоподобия:

$$\ln L(\lambda) = -\ln(x_1!\, x_2! \ldots x_n!) + \left(\sum_{i=1}^{n} x_i\right) \ln \lambda - \lambda n.$$

Используя необходимые условия экстремума $L(\lambda)$, получим уравнение для определения оценки параметра λ:

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0.$$

Отсюда следует, что $\lambda^* = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$.

2.4. Интервальные оценки. Доверительные интервалы

При статистической обработке результатов эксперимента часто необходимо не только найти оценку ϑ^* неизвестного параметра ϑ, но и охарактеризовать точность этой оценки. С этой целью вводят понятие доверительного интервала.

Доверительным интервалом неизвестного параметра ϑ называется интервал $\left(\vartheta^* - \varepsilon;\ \vartheta^* + \varepsilon\right)$, содержащий (накрывающий) истинное значение неизвестного параметра ϑ с заданной вероятностью $\gamma = 1 - \alpha$, то есть

$$P\left(\vartheta^* - \varepsilon < \vartheta < \vartheta^* + \varepsilon\right) = \gamma = 1 - \alpha .$$

Значение α — *уровень значимости*, число γ называется *доверительной вероятностью* и характеризует *надежность* оценки.

Величина ε определяет *точность* интервального оценивания, зависит от объема выборки n, доверительной вероятности γ и показывает *предельную ошибку* выборки. При увеличении объема выборки предельная ошибка уменьшается, а с приближением доверительной вероятности к единице — увеличивается. Выбор доверительной вероятности определяется конкретными условиями. Обычно используются значения γ, равные 0,90; 0,95; 0,99.

Задача 6

Найти доверительный интервал для неизвестного параметра ϑ по заданной доверительной вероятности γ.

Решение

1. Пусть $\{x_1, x_2, ..., x_n\}$ — выборка наблюдений случайной величины X. Отметим, что границы доверительного интервала $\vartheta_1 = \vartheta^* - \varepsilon$ и $\vartheta_2 = \vartheta^* + \varepsilon$ являются функциями выборки $\vartheta_1 = \vartheta_1\left(x_1, x_2, ..., x_n\right)$, $\vartheta_2 = \vartheta_2\left(x_1, x_2, ..., x_n\right)$.

2. Вместо случайной величины X рассмотрим *другую* вспомогательную случайную величину $Y = Y\left(\vartheta, \vartheta^*\right)$, распределение которой известно и *не содержит параметров*. Функция $Y = Y\left(\vartheta, \vartheta^*\right)$ непрерывна и строго монотонна по ϑ.

3. Рассмотрим равенство $P\left(\left|Y\left(\vartheta, \vartheta^*\right)\right| < u_\gamma\right) = \gamma$, где γ — заданная доверительная вероятность, u_γ — квантиль распределения случайной величины Y.

Левую часть расписываем по закону распределения вспомогательной случайной величины Y.

4. Из условия 3 находим квантиль u_γ распределения случайной величины Y.

5. Решая неравенство $\left|Y\left(\vartheta, \vartheta^*\right)\right| < u_\gamma$ относительно ϑ, найдем границы ϑ_1 и ϑ_2 доверительного интервала для ϑ, следовательно, выполняется неравенство $\vartheta_1 < \vartheta < \vartheta_2$.

Непараметрические распределения

Стандартное распределение — это нормальное распределение с параметрами: $a = 0,\ \sigma^2 = 1$.

Плотность распределения $f(x)$ определяется формулой

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

На рис. 2.2 приведен график функции $f(x)$ и u_γ — симметричная *квантиль стандартного распределения порядка* γ.

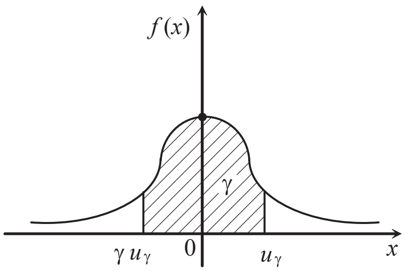


Рис. 2.2. График функции плотности стандартного распределения

Распределение Хи-квадрат с k степенями свободы — распределение случайной величины $\chi^2(k)$, которая равна сумме квадратов k независимых случайных величин U_i, $i = 1, 2, \ldots, k$, каждая из которых распределена стандартно:

$$\chi^2(k) = U_1^2 + U_2^2 + \ldots + U_k^2.$$

Плотность распределения $f_{\chi^2}(x)$ определяется формулой

$$f_{\chi^2}(x) = \begin{cases} 0, & x \le 0, \\ \dfrac{1}{2^{k/2}\Gamma\left(\dfrac{k}{2}\right)} x^{(k-2)/2} e^{-x/2}, & x > 0, \end{cases}$$

где $\Gamma(z) = \int\limits_0^\infty t^{z-1} e^{-t} dt$ — гамма-функция; z — любое комплексное число, кроме $0,\ -1,\ -2,\ \ldots$.

На рис. 2.3 приведены графики функции $f_{\chi^2}(x)$ с k_1 и k_2 степенями свободы и χ_γ^2 — *квантиль распределения Хи-квадрат* с *k степенями свободы порядка* γ.

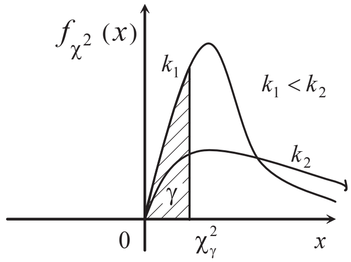


Рис. 2.3. График функции плотности распределения $\chi^2(k)$

Распределение Стьюдента с *k степенями свободы* — распределение случайной величины $T(k)$, равной отношению двух независимых случайных величин:

$$T(k) = \frac{U}{\sqrt{\chi^2(k)/k}},$$

где U — распределена стандартно.

Плотность распределения $f_T(x)$ определяется формулой:

$$f_T(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}} \left(1 + \frac{x^2}{k}\right)^{-(k+1)/2}, \quad -\infty < x < +\infty.$$

На рис. 2.4 приведены графики функции $f_T(x)$ с k_1 и k_2 степенями свободы и t_γ — симметричная *квантиль распределения Стьюдента* с *k* степенями свободы *порядка* γ.

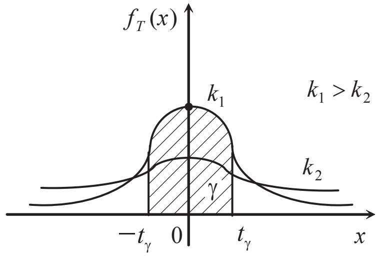


Рис. 2.4. График функции плотности распределения Стьюдента

Замечание. Все упоминавшиеся квантили находятся либо по таблицам, либо с помощью компьютерных вычислений.

2.5. Доверительные интервалы для параметров нормально распределенной генеральной совокупности

2.5.1. Доверительный интервал для математического ожидания случайной величины при известной дисперсии

Построим доверительный интервал для математического ожидания случайной величины с нормальным распределением при известной дисперсии σ^2 и доверительной вероятности γ.

1. Пусть $\{x_1, x_2, ..., x_n\}$ — выборка из нормально распределенной генеральной совокупности.

2. В качестве оценки математического ожидания a возьмем выборочное среднее:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Выборочное среднее $\bar{x}$ в данном случае имеет нормальное распределение с числовыми характеристиками:

$$M(\bar{x}) = a , \ D(\bar{x}) = \frac{1}{n}\sigma^2 .$$

3. Рассмотрим вспомогательную случайную величину $Y = \dfrac{\bar{x} - a}{\sigma / \sqrt{n}}$.

Она подобрана так, чтобы ее распределение было стандартным независимо от параметра a:

$$M(Y) = 0 , \ D(Y) = 1 .$$

Действительно, так как $f(\bar{x}) = \dfrac{1}{\sqrt{2\pi}\,\sigma / \sqrt{n}} e^{-\frac{(\bar{x}-a)^2}{2\sigma^2/n}}$ и $Y = Y(\bar{x})$ — функция от случайной величины $\bar{x}$, то

$$f(y) = f(\bar{x}(y)) \cdot |\bar{x}'(y)| .$$

Поскольку $\bar{x} = y\sigma / \sqrt{n} + a$, то $\bar{x}'(y) = \sigma / \sqrt{n}$.

Следовательно,

$$f(y) = \sigma / \sqrt{n} \cdot \frac{1}{\sqrt{2\pi}\,\sigma / \sqrt{n}} e^{-\frac{\left(y\sigma/\sqrt{n}+a-a\right)^2}{2\sigma^2/n}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2\sigma^2/n}{2\sigma^2/n}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$$

Таким образом, действительно, вспомогательная случайная величина Y распределена стандартно.

4. По центральной предельной теореме [4]:

$$P\left(|Y| < u_\gamma\right) = 2\Phi\left(u_\gamma\right) \Rightarrow 2\Phi\left(u_\gamma\right) = \gamma.$$

5. Найдем из этого равенства квантиль u_γ (по таблице или с помощью компьютерных вычислений).

6. Решая неравенство $-u_\gamma < \dfrac{\bar{x}-a}{\sigma/\sqrt{n}} < u_\gamma$ относительно a, получим, что с вероятностью γ выполняется следующее условие:

$$\bar{x} - u_\gamma \cdot \frac{\sigma}{\sqrt{n}} < a < \bar{x} + u_\gamma \cdot \frac{\sigma}{\sqrt{n}}.$$

Искомый доверительный интервал для математического ожидания a:

$$\left(\bar{x} - u_\gamma \cdot \frac{\sigma}{\sqrt{n}};\quad \bar{x} + u_\gamma \cdot \frac{\sigma}{\sqrt{n}}\right).$$

2.5.2. Доверительный интервал для математического ожидания случайной величины при неизвестной дисперсии и малой выборки

Определим доверительный интервал для математического ожидания нормально распределенной случайной величины при неизвестной дисперсии и *малой выборке* (объем выборки $n < 30$) по доверительной вероятности γ.

Для построения доверительного интервала используем распределение Стьюдента с $(n-1)$ степенями свободы:

$$f_T(x, n-1) = \frac{\Gamma\left(\frac{n}{2}\right)\left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi(n-1)}}.$$

Теорема 1. *Если элементы выборки* $\{x_1, x_2, \ldots, x_n\}$ *независимы и каждый из них распределяется нормально с параметрами* (a, σ^2), *то случайная величина*

$$Y = \frac{\bar{x} - a}{\sqrt{D_{\text{В}}}}\sqrt{n-1}$$

имеет распределение Стьюдента с $(n-1)$ *степенями свободы.* Следовательно, из теоремы следует, что

$$P(|Y| < t_\gamma) = 2\int_0^{t_\gamma} f_T(x, n-1)\,dx .$$

Итак, для нахождения границ доверительного интервала для математического ожидания a, отвечающего доверительной вероятности γ, введем вспомогательную случайную величину $Y = \frac{\bar{x} - a}{\sqrt{D_{\text{В}}}}\sqrt{n-1}$, где $D_{\text{В}}$ находится по выборке $\{x_1, x_2, \ldots, x_n\}$. С одной стороны, имеем $P(|Y| < t_\gamma) = 2\int_0^{t_\gamma} f_T(x, n-1)\,dx$, с другой —

$$P(|Y| < t_\gamma) = P\left(\left|\frac{\bar{x} - a}{\sqrt{D_{\text{В}}}}\sqrt{n-1}\right| < t_\gamma\right) = P\left(\bar{x} - t_\gamma\sqrt{\frac{D_{\text{В}}}{n-1}} < a < \bar{x} + t_\gamma\sqrt{\frac{D_{\text{В}}}{n-1}}\right) = \gamma .$$

Из условия $P(|Y| < t_\gamma) = \gamma$ находим квантиль t_γ.

Искомый интервал для математического ожидания a получен:

$$\left(\bar{x} - t_\gamma\sqrt{\frac{D_{\text{В}}}{n-1}};\quad \bar{x} + t_\gamma\sqrt{\frac{D_{\text{В}}}{n-1}}\right).$$

Задача 7

На контрольных испытаниях случайно выбранных из партии 15 осветительных ламп были определены средняя продолжительность работы лампы $\bar{x} = 3000$ ч и среднеквадратическое отклонение $\sigma \ = 20$ ч. Найти с вероятностью $\gamma = 0,99$ доверительный интервал для средней продолжительности работы лампы в целом для партии.

Решение

Поскольку объем выборки $n = 15$, что меньше 30, то выборка — малая. Доверительный интервал для средней продолжительности службы лампы во всей партии (генеральная средняя — математическое ожидание a) определяется неравенством

$$\bar{x} - t_\gamma \sqrt{\frac{\sigma_{\text{В}}^2}{n-1}} < a < \bar{x} + t_\gamma \sqrt{\frac{\sigma_{\text{В}}^2}{n-1}}.$$

По таблице распределения Стьюдента (прил. 2): при γ = 0,99, n–1 = = 15–1 = 14 имеем квантиль $t_\gamma = 2,98$.

Вычислим: $\sqrt{\frac{\sigma_{\text{В}}^2}{n-1}} = \sqrt{\frac{20^2}{14}} = 5,35$. Тогда получим:

$$t_\gamma \sqrt{\frac{\sigma_{\text{В}}^2}{n-1}} = 2,98 \cdot 5,35 = 15,94.$$

Итак, доверительный интервал для математического ожидания a:

$$3000 - 15,94 < a < 3000 + 15,94,$$

или

$$2984,06 < a < 3015,94.$$

Таким образом, с вероятностью $0,99$ можно утверждать, что средняя продолжительность службы лампы для всей партии колеблется в пределах от $2984,06$ до $3015,94$ ч.

2.5.3. Доверительный интервал для дисперсии случайной величины с неизвестным математическим ожиданием

Построим доверительный интервал для дисперсии случайной величины с нормальным распределением при неизвестном математическом ожидании.

Задача 8

Определить доверительный интервал для дисперсии нормально распределенной случайной величины с неизвестным математическим ожиданием по доверительной вероятности γ.

Задача решается с помощью распределения χ^2 с $(n-1)$ степенями свободы:

$$f_{\chi^2}(x,n-1)=\frac{1}{2^{(n-1)/2}\Gamma\left(\frac{n-1}{2}\right)}x^{(n-3)/2}e^{-x/2}.$$

Теорема 2. *Если элементы выборки $\{x_1, x_2, ..., x_n\}$ независимы и каждый из них распределяется нормально с параметрами (a,σ^2), то $\bar{x}$ и $D_{\text{В}}$ — независимы, причем $\bar{x}$ распределено нормально с параметрами $(a,\sigma^{2/n})$, случайная величина $Y=\frac{nD_{\text{В}}}{\sigma^2}$ имеет распределение χ^2 с $(n-1)$ степенями свободы.*

По теореме 2 $P(\chi_1^2<Y<\chi_2^2)=\gamma$, величины χ_1^2 и χ_2^2 находятся из уравнений $\chi_1^2: \int\limits_{\chi_1^2}^{+\infty} f_{\chi^2}(x,n-1)dx=\frac{1+\gamma}{2}$, $\chi_2^2: \int\limits_{\chi_2^2}^{+\infty} f_{\chi^2}(x,n-1)dx=\frac{1-\gamma}{2}$.

После подстановки $Y=\frac{nD_{\text{В}}}{\sigma^2}$ в неравенство $\chi_1^2<Y<\chi_2^2$ получаем

$$\frac{nD_{\text{В}}}{\chi_2^2}<\sigma^2<\frac{nD_{\text{В}}}{\chi_1^2}.$$

Искомый интервал для математического ожидания σ^2:

$$\left(\frac{nD_{\text{В}}}{\chi_2^2};\ \frac{nD_{\text{В}}}{\chi_1^2}\right).$$

2.6. Доверительный интервал для вероятности успеха в схеме Бернулли и параметра λ распределения Пуассона

Определим доверительный интервал для вероятности успеха в схеме Бернулли.

Задача 9

Пусть в *n* независимых испытаниях успех наступил *x* раз. Найти доверительный интервал для вероятности *p* успеха в одном испытании по доверительной вероятности γ.

Решение

Рассмотрим случай, когда n — велико. Тогда эффективной оценкой вероятности p успеха в одном испытании является относительная частота успеха $p^* = w = \dfrac{x}{n}$.

По теореме Муавра-Лапласа относительная частота w имеет нормальное распределение с параметрами $\left(a, \sigma^2\right)$: $a = p$, $\sigma^2 = pq/n$, где $q = 1 - p$.

Рассмотрим вспомогательную случайную величину $Y = \dfrac{w-p}{\sqrt{pq/n}}$, которая, следовательно, имеет стандартное распределение независимо от значения p. При больших n тогда имеем:

$$P\left(\left|\frac{w-p}{\sqrt{pq/n}}\right| < u_\gamma\right) \approx 2\Phi\left(u_\gamma\right) \Rightarrow 2\Phi\left(u_\gamma\right) \approx \gamma .$$

Найдем из этого приближенного равенства квантиль u_γ (прил. 1 или с помощью компьютерных вычислений).

Решая неравенство $-u_\gamma < \dfrac{w-p}{\sqrt{pq/n}} < u_\gamma$ относительно p, получим, что с вероятностью $\approx \gamma$ выполняется следующее условие:

$$w - u_\gamma \sqrt{\frac{pq}{n}} < p < w + u_\gamma \sqrt{\frac{pq}{n}} .$$

Заменяя значения p и q в левой и правой частях этого неравенства их оценками $p^* = w$ и $q^* = 1 - w$, получаем, что доверительный интервал для вероятности успеха в схеме Бернулли приближенно имеет вид

$$w - u_\gamma \sqrt{\frac{w(1-w)}{n}} < p < w + u_\gamma \sqrt{\frac{w(1-w)}{n}} .$$

Задача 10

Пусть $\{x_1, x_2, ..., x_n\}$ — выборка из генеральной совокупности, имеющей распределение Пуассона с неизвестным параметром λ:

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda},$$

где x принимает неотрицательные целочисленные значения $x = 0, 1, 2, \dots$.

Найти доверительный интервал для параметра λ по доверительной вероятности γ.

Решение

Рассмотрим случай, когда n — велико. Тогда доверительный интервал для параметра λ приближенно имеет вид

$$\bar{x} - u_\gamma \sqrt{\frac{\bar{x}}{n}} < \lambda < \bar{x} + u_\gamma \sqrt{\frac{\bar{x}}{n}},$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ — выборочное среднее; u_γ — квантиль стандартного распределения.

2.7. Статистические оценки параметров распределения

В данном разделе приведена методика проведения лабораторной работы по решению следующих типов задач статистического оценивания:

- по сделанной выборке объема n и заданной точности $\varepsilon > 0$ вычислить доверительную вероятность γ, с которой интервал $(\vartheta^* - \varepsilon; \vartheta^* + \varepsilon)$ накроет неизвестную генеральную характеристику ϑ;
- по сделанной выборке объема n и с заданной надежностью γ найти доверительный интервал для неизвестной характеристики ϑ:

$$(\vartheta^* - \varepsilon < \vartheta < \vartheta^* + \varepsilon);$$

- определить минимальный объем выборки n, при котором с заданной надежностью γ обеспечивается указанная предельная ошибка ε.

Цель работы — научиться решать основные типы задач статистического оценивания.

Задача 11

В результате 5%-ного собственно-случайного бесповторного обследования в исправительной колонии получены следующие данные о распределении осужденных по срокам лишения свободы.

Сроки лишения свободы, лет	До 2	2–4	4–6	6–8	Более 8
Число осужденных, чел.	19	17	11	8	5

Найти:

а) с вероятностью 0,95 границы для среднего срока лишения свободы по колонии в целом;

б) доверительную вероятность того, что средний срок лишения свободы по колонии в целом отличается от среднего в выборке не более чем на 0,5 года;

в) границы, в которых с вероятностью 0,95 заключена доля осужденных со сроком лишения свободы более 6 лет;

г) минимальный объем выборки, гарантирующий с вероятностью 0,97 предельную ошибку выборки 0,3 года.

Порядок выполнения работы

Открыть **Microsoft Excel**, в новом документе удалить листы 2 и 3, сохранить файл в своей папке под именем «Задача 2». Далее также периодически сохранять документ. Переименовать *Лист1* в *Расчетная таблица*. Образец расчетной таблицы приведен на рис. 2.5.

Заполнение расчетной таблицы

Используя приведенный образец, заполнить таблицу в следующем порядке:

- набрать строку заголовков, сделать в ней шрифт полужирным, выравнивание по центру. Заполнить столбцы A, B, и D данными из условия задачи, в строке *Итого* в столбце D получить результат с помощью «*Автосуммы*»;
- вычислить значения столбцов E, F, подсчитайть итоговые суммы в этих столбцах;
- ниже расчетной таблицы сделать подписи и вычислить длину интервала h, выборочное среднее $x_{\text{ср}}$, $x^2{}_{\text{ср}}$, выборочную дисперсию $D_{\text{В}}$ и выборочное среднее квадратическое *отклонение* SI;

- так как в пунктах *а*, *б*, *в*, *г* задачи некоторые величины принимают разные значения, оформить нижнюю часть таблицы следующим образом: слева набрать подписи величин, а сверху указать пункты. Ответы на вопросы задачи расположить правее расчетной таблицы (рис. 2.6).

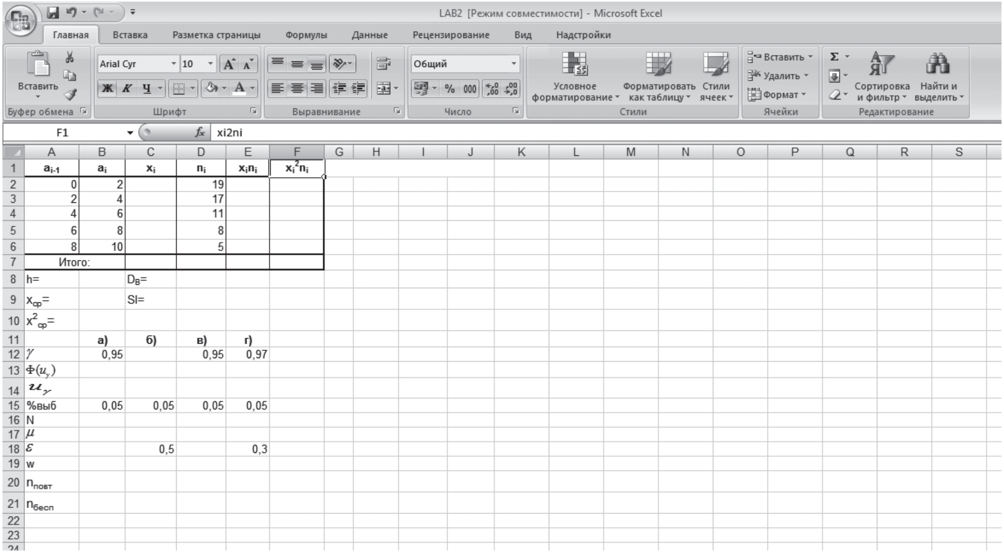


Рис. 2.5. Расчетная таблица

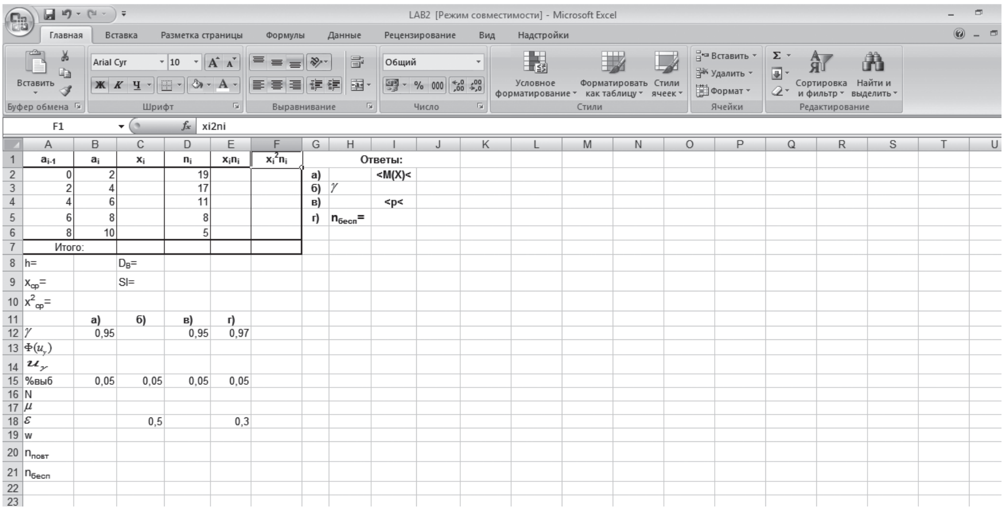


Рис. 2.6. Расчетная таблица с ответами

Описание процесса решения задачи по пунктам

Решение пункта *а* задачи 11

Занести в соответствующую ячейку (B12) данное в условии значение доверительной вероятности γ: $\gamma = 0{,}95$.

Вычислить значение функции Лапласа $\Phi(u_\gamma)$:

$$\Phi(u_\gamma) = \frac{\gamma}{2},$$

то есть в ячейку B13 ввести следующую формулу:

$$= B12 / 2.$$

Зная значение функции $\Phi(u_\gamma)$, найти квантиль u_γ по таблице значений функции Лапласа (прил. 1) и записать его в ячейку B14.

Процент выборки, данный в условии, перевести в доли единицы и занести в ячейку B15.

Объем генеральной совокупности N вычисляется следующим образом:

$$N = \frac{\sum n_i}{\%_{\text{выб}}}.$$

В ячейку B16 записать следующую формулу:

$$= D7 / B15.$$

Средняя ошибка выборки μ вычисляется по формуле

$$\mu = \sqrt{\frac{D}{\sum n_i}\left(1 - \frac{\sum n_i}{N}\right)}.$$

В ячейку B17 записать следующую формулу:

$$= \text{КОРЕНЬ}\Big(D8 / D7 * \big(1 - D7 / B16\big)\Big).$$

Предельная ошибка выборки ε вычисляется по формуле

$$\varepsilon = u_\gamma \mu.$$

В ячейку B18 записать следующую формулу:

$$= B14 * B17 .$$

В ячейке I2 сделать надпись: $< M(X) <$.

α и β — границы доверительного интервала для генерального среднего $M(X)$ вычисляются соответственно в ячейках H2 и J2 по формулам

$$\alpha = x_{\text{ср}} - \varepsilon ,$$

$$\beta = x_{\text{ср}} + \varepsilon ,$$

то есть в ячейку H2 введите формулу

$$= B9 - B18 ,$$

в ячейку J2 — формулу

$$= B9 + B18 .$$

Решение пункта *б* задачи 11

Теперь предельная ошибка выборки ε дана — записать ее значение 0,5 в ячейку C18.

Средняя ошибка выборки μ не изменится, так как рассчитывается по той же формуле, в которой все величины прежние. Поэтому необходимо скопировать численное значение ячейки B17 в ячейку C17.

Вычислить значение квантили u_γ по формуле

$$u_\gamma = \frac{\varepsilon}{\mu} .$$

Для этого в ячейку C14 ввести следующую формулу:

$$= C18 / C17 .$$

Зная значение квантили u_γ , по таблице значений функции Лапласа (прил. 1) найти $\Phi(u_\gamma)$ и записать его в ячейку C13.

Далее, определить доверительную вероятность γ :

$$\gamma = 2\Phi(u_\gamma) .$$

В ячейку C12 записать следующую формулу:

$$= 2 * C13 .$$

Перенести последний результат в ответы (ячейка I3).

Решение пункта *в* задачи 11

Занести в соответствующую ячейку (D12) данное в условии значение доверительной вероятности γ: $\gamma = 0,95$.

Вычислить значение функции Лапласа $\Phi(u_\gamma)$:

$$\Phi(u_\gamma) = \frac{\gamma}{2},$$

то есть в ячейку D13 ввести следующую формулу:

$$= D12 / 2.$$

Зная $\Phi(u_\gamma)$, найти квантиль u_γ по таблице значений функции Лапласа (прил. 1) и записать его в ячейку D14.

Процент выборки, данный в условии, перевести в доли единицы и занести в ячейку D15.

Объем генеральной совокупности N вычисляется следующим образом:

$$N = \frac{\sum n_i}{\%_{\text{выб}}}.$$

В ячейку D16 записать следующую формулу:

$$= D7 / D15.$$

Выборочная доля w заключенных со сроком лишения свободы более 6 лет в выборке согласно условию задачи:

$$w = \frac{8+5}{\sum n_i},$$

то есть в ячейку D19 ввести формулу

$$= (D5 + D6) / D7.$$

Поскольку оценивается теперь доля, то средняя ошибка доли выборки μ теперь вычисляется по новой формуле:

$$\mu = \sqrt{\frac{w(1-w)}{\sum n_i}\left(1 - \frac{\sum n_i}{N}\right)}.$$

В ячейку D17 ввести следующую формулу:

$$= \text{КОРЕНЬ}\left(D19*(1-D19)/D7*\left(1-D7/D16\right)\right).$$

Подсчитать предельную ошибку выборки ε:

$$\varepsilon = u_\gamma \mu .$$

В ячейку D18 записать следующую формулу:

$$= D14*D17 .$$

В ячейке I4 сделать надпись: $< p <$.

α и β — границы для генеральной доли вычисляются соответственно в ячейках H4 и J4 по формулам

$$\alpha = w - \varepsilon ,$$
$$\beta = w + \varepsilon ,$$

то есть в ячейку H4 ввести формулу

$$= D19 - D18 ,$$

в ячейку J4 — формулу

$$D19 + D18 .$$

Решение пункта *г* задачи 11

Занести в соответствующую ячейку (E12) данное в условии значение доверительной вероятности γ: $\gamma = 0{,}97$.

Вычислить значение функции Лапласа $\Phi(u_\gamma)$:

$$\Phi(u_\gamma) = \frac{\gamma}{2} ,$$

то есть в ячейку E13 ввести следующую формулу:

$$= E12/2 .$$

Зная значение функции $\Phi(u_\gamma)$, найти квантиль u_γ по таблице значений функции Лапласа (прил. 1) и записать его в ячейку E14.

Процент выборки, данный в условии, перевести в доли единицы и занести в ячейку E15.

Объем генеральной совокупности N вычисляется следующим образом:

$$N = \frac{\sum n_i}{\%_{\text{выб}}}.$$

В ячейку E16 записать следующую формулу:

$$= D7 / E15.$$

Предельная ошибка выборки ε дана, записать ее значение 0,3 в ячейку E18.

Вычислить минимальный объем повторной выборки по формуле

$$n_{\text{повт}} = D_{\text{В}} \frac{u_\gamma^2}{\varepsilon^2}.$$

В ячейку E20 записать следующую формулу:

$$= D8*(E14 / E18)\hat{}2.$$

Минимальный объем бесповторной выборки:

$$n_{\text{бесп}} = \frac{n_{\text{повт}} N}{n_{\text{повт}} + N}.$$

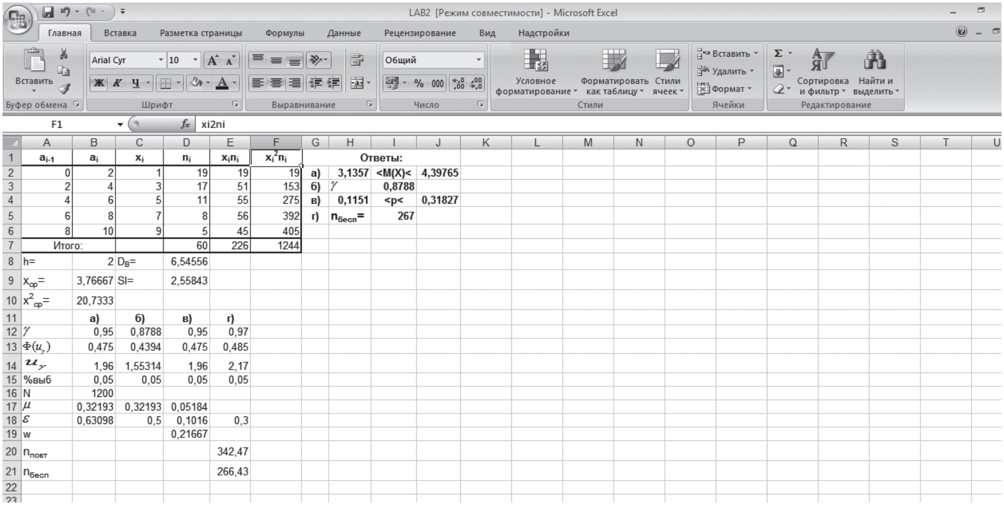
В ячейку E21 записать следующую формулу:

$$= E20 * E16 / (E20 + E16).$$

Округлить полученный результат до целых в большую сторону и оформить ответ пункта *г*.

Добавить линии в расчетной таблице. Заголовки *а*, *б*, *в*, *г* и ответы сделать полужирным шрифтом.

На рис. 2.7 изображена итоговая расчетная таблица, которая должна получиться.



	A	B	C	D	E	F	G	H	I	J
1	a_{i-1}	a_i	x_i	n_i	x_in_i	$x_i^2n_i$		Ответы:		
2	0	2	1	19	19	19	а)	3,1357	<M(X)<	4,39765
3	2	4	3	17	51	153	б)	γ	0,8788	
4	4	6	5	11	55	275	в)	0,1151	<p<	0,31827
5	6	8	7	8	56	392	г)	$n_{бесп}$=	267	
6	8	10	9	5	45	405				
7	Итого:			60	226	1244				
8	h=	2	$D_В$=	6,54556						
9	$x_{ср}$=	3,76667	Sl=	2,55843						
10	$x^2_{ср}$=	20,7333								
11		а)	б)	в)	г)					
12	γ	0,95	0,8788	0,95	0,97					
13	$\Phi(u_\gamma)$	0,475	0,4394	0,475	0,485					
14	u_γ	1,96	1,55314	1,96	2,17					
15	%выб	0,05	0,05	0,05	0,05					
16	N	1200								
17	μ	0,32193	0,32193	0,05184						
18	ε	0,63098	0,5	0,1016	0,3					
19	w			0,21667						
20	$n_{повт}$				342,47					
21	$n_{бесп}$				266,43					

Рис. 2.7. Итоговая расчетная таблица

Индивидуальное задание

Создать лист *Расчетная таблица №*. Решить свой вариант задания.

Результаты работы

В результате выполненной работы студент должен продемонстрировать преподавателю готовый файл *Задача 2.xlsx*, содержащий 3 листа:

- лист 1 (*Титульный лист*) — титульный лист к работе, на котором указаны название работы, номер варианта, Ф. И.О. студента, номер группы, дата выполнения работы;
- лист 2 (*Расчетная таблица*) — таблица исходных данных и решения задачи;
- лист 3 (*Расчетная таблица №...*) — таблица исходных данных и решения задания своего варианта.

Варианты индивидуального задания

Вариант 1

На ткацкой фабрике из 1000 ткачих произведена собственно-случайная бесповторная выборка 100 человек. В результате получены следующие данные о распределении ткачих по уровню дневной выработки.

Уровень дневной выработки, м	30–40	40–50	50–60	60–70
Число ткачих, чел.	30	33	24	13

Вычислить:
а) с вероятностью 0,954 границы для средней дневной выработки одной ткачихи по фабрике в целом;
б) границы, в которых с вероятностью 0,997 заключена доля ткачих с дневной выработкой более 60 м;
в) минимальный объем выборки, гарантирующий с вероятностью 0,98 предельную ошибку выборки 2 м.

Вариант 2

Организована собственно-случайная 10%-ная бесповторная выборка по предприятию. Получены следующие данные о распределении рабочих по заработной плате.

Зарплата, тыс. руб.	20–30	30–40	40–50	50–60	60–70	70–80
Число рабочих, чел.	8	19	28	32	42	21

Найти:
а) с вероятностью 0,9964 границы для средней зарплаты рабочих завода;
б) доверительную вероятность того, что средняя зарплата рабочих завода отличается от средней в выборке не более чем на 0,7 тыс. руб.;
в) минимальный объем выборки, гарантирующий с вероятностью 0,97 предельную ошибку выборки 0,3 тыс. руб.

Вариант 3

Имеются следующие данные выборки о распределении рабочих завода по стажу работы (применялся собственно-случайный бесповторный 5%-ный отбор).

Стаж работы, лет	0–5	5–10	10–15	15–20	20–25	25–30
Число рабочих, чел.	20	30	40	50	40	20

Найти:
а) с вероятностью 0,95 границы для среднего стажа работы по заводу в целом;
б) доверительную вероятность того, что средний срок стажа по заводу в целом отличается от среднего в выборке не более чем на 0,6 года;
в) границы, в которых с вероятностью 0,95 заключена доля рабочих со стажем работы более 20 лет.

Вариант 4

В результате собственно-случайного бесповторного 5%-ного отбора, проведенного с целью обследовать жилищные условия жителей города, получены следующие данные:

Общая площадь на 1 чел., м^2	До 5	5—10	10—15	15—20	20—25	25—30	30 и более
Число жителей, чел.	8	95	204	270	210	130	83

Найти:

а) с вероятностью 0,954 границы для среднего размера общей площади, приходящейся на 1 человека по городу в целом;

б) границы, в которых с вероятностью 0,954 заключена доля лиц, имеющих менее 10 м^2 на человека;

в) минимальный объем выборки, гарантирующий с вероятностью 0,97 предельную ошибку выборки 0,7 м^2.

Вариант 5

Для установления средней продолжительности остаточного срока заключения в исправительной колонии произведена 5%-ная собственно-случайная бесповторная выборка. Данные выборочного обследования представлены ниже.

Остаточный срок заключения, лет	До 1	1—2	2—3	3—4	4—5	5—6	Более 6
Число заключенных, чел.	11	19	39	34	15	5	2

Найти:

а) с вероятностью 0,95 границы для средней продолжительности остаточного срока заключения по колонии в целом;

б) границы, в которых с вероятностью 0,95 заключена доля заключенных колонии с остаточным сроком заключения менее 2 лет;

в) необходимый объем выборки при определении средней продолжительности остаточного срока заключения, гарантирующий с вероятностью 0,99 предельную ошибку выборки 0,5 года.

Вариант 6

Выборочным путем получены следующие данные об урожайности ржи в области (применялся собственно-случайный бесповторный 10%-ный отбор):

Урожайность, ц/га	12—14	14—16	16—18	18—20	20—22	22—24	24—26
Площадь, га	36	114	218	272	166	132	62

Найти:

а) с вероятностью 0,9545 границы для средней урожайности на всей площади области, занятой рожью;
б) вероятность того, что средняя урожайность, полученная в выборке, отличается от средней урожайности на всей площади области, занятой рожью, не более чем на 0,2 ц/га;
в) границы, в которых с вероятностью 0,996 заключена доля площадей с урожайностью менее 22 ц/га в области.

Вариант 7

Получены следующие данные о распределении продавцов магазина по выработке (применялся собственно-случайный бесповторный 3%-ный отбор):

Выработка продавцов, тыс. руб.	80–100	100–120	120–140	140–160	160–180
Число продавцов, чел.	5	10	20	10	5

Вычислить:

а) с вероятностью 0,954 границы для средней выработки одного продавца по магазину в целом;
б) границы, в которых с вероятностью 0,997 заключена доля продавцов с выработкой более 160 тыс. руб.;
в) минимальный объем выборки, гарантирующий с вероятностью 0,98 предельную ошибку выборки 1 тыс. руб.

Вариант 8

Имеются результаты измерения роста случайно отобранных 100 студентов (применялся собственно-случайный бесповторный 5%-ный отбор).

Рост, см	158–162	162–166	166–170	170–174	174–178	178–182	182–186
Число студентов, чел.	10	14	26	28	12	8	2

Найти:

а) с вероятностью 0,95 границы для среднего роста студентов по городу в целом;
б) границы, в которых с вероятностью 0,95 заключена доля студентов с ростом более 174 см;
в) необходимый объем выборки при определении среднего роста студентов, гарантирующий с вероятностью 0,99 предельную ошибку выборки 6 см.

Вариант 9

Получено следующее распределение выборки (применялся собственно-случайный бесповторный 3%-ный отбор):

Частичный интервал	1–5	5–9	9–13	13–17	17–21
Частота вариант интервала	10	20	50	12	8

Вычислить:

а) с вероятностью 0,954 границы для генерального среднего;

б) доверительную вероятность того, что выборочное среднее отличается от генерального среднего не более чем на 0,7;

в) минимальный объем выборки, гарантирующий с вероятностью 0,98 предельную ошибку выборки 1.

Вариант 10

Получено следующее распределение выборки (применялся собственно-случайный бесповторный 5%-ный отбор):

Частичный интервал	10–15	15–20	20–25	25–30	30–35
Частота вариант интервала	2	4	8	4	2

Вычислить:

а) с вероятностью 0,95 границы для генерального среднего;

б) доверительную вероятность того, что выборочное среднее отличается от генерального среднего не более чем на 0,3;

в) минимальный объем выборки, гарантирующий с вероятностью 0,99 предельную ошибку выборки 0,5.

3. Корреляционная зависимость

Одной из важнейших задач статистики является анализ зависимостей между изучаемыми признаками, включающий в себя:

- установление существующей взаимосвязи;
- определение роли признаков в данной взаимосвязи (*факторного* признака, оказывающего влияние на другие связанные с ним признаки, или *результирующего* признака, изменяющегося под действием факторного признака);
- оценку тесноты связи (количественной меры зависимости признаков);
- выбор аналитической формы изучаемой зависимости в виде определенной математической функции (*регрессии*) и оценку ее параметров;
- проверку адекватности выбранного функционального вида зависимости (статистической модели взаимосвязи).

Для многих явлений характерно, что существующие причинно-следственные связи «размыты» действием многих случайных факторов. Поэтому в исследуемой взаимосвязи при одном и том же значении факторного признака (переменная X) могут наблюдаться различные значения результативного признака (переменная Y). Среднюю этого множества значений переменной Y, отвечающих определенному значению x переменной X, называют условной средней признака Y при $X = x$ и обозначают $\bar{y}(x)$. Зависимость, при которой изменению значений факторного признака соответствует изменение среднего значения результативного признака, называется *корреляционной*.

Полученные данные выборочного наблюдения, проведенного в целях установления зависимости между признаками Y и X, группируют и представляют в виде *корреляционной таблицы*.

Корреляционная таблица

$X \backslash Y$	y_1	y_2	…	y_s	Всего	$\bar{y}_i$
x_1	n_{11}	n_{12}	…	n_{1s}	n_{x1}	$\bar{y}_1$
x_2	n_{21}	n_{22}	…	n_{2s}	n_{x2}	$\bar{y}_2$
…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{ks}	n_{xk}	$\bar{y}_k$
Всего	n_{y1}	n_{y2}	…	n_{ys}	$n = \sum_{i=1}^{k} n_{x_i} = \sum_{j=1}^{s} n_{y_j}$	

Таблица включает в себя следующие элементы:

- $x_1, x_2, \ldots, x_k$ — варианты признака X (для дискретного ряда) или середины интервалов (для интервального ряда);
- $y_1, y_2, \ldots, y_s$ — варианты признака Y (для дискретного ряда) или середины интервалов (для интервального ряда);
- n_{ij} — частоты совместного появления значений признаков X и Y, попавших в i-ю группу по X и в j-ю группу по Y, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, s$;
- $n_{x_1}, n_{x_2}, \ldots, n_{x_k}$ — частоты значений $x_1, x_2, \ldots, x_k$ признака X, вычисляемые как суммы частот n_{ij} по каждой строке:

$$n_{x_1} = \sum_{j=1}^{s} n_{1j}\ ,\ n_{x_2} = \sum_{j=1}^{s} n_{2j}\ ,\ \ldots,\ n_{x_k} = \sum_{j=1}^{s} n_{kj}\ ;$$

- $n_{y_1}, n_{y_2}, \ldots, n_{y_s}$ — частоты значений $y_1, y_2, \ldots, y_s$ признака Y, вычисляемые как суммы частот n_{ij} по каждому столбцу:

$$n_{y_1} = \sum_{i=1}^{k} n_{i1}\ ,\ n_{y_2} = \sum_{i=1}^{k} n_{i2}\ ,\ \ldots,\ n_{y_s} = \sum_{i=1}^{k} n_{is}\ ;$$

- $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ — условные средние признака Y, вычисляемые как групповые средние значений $y_1, y_2, \ldots, y_s$ по каждой группе значений признака X, то есть $\bar{y}_i$ — средняя признака Y при условии $X = x_i$, $i = 1, 2, \ldots, k$:

$$\bar{y}_1 = \frac{1}{n_{x_1}}\left(y_1 n_{11} + y_2 n_{12} + \ldots + y_s n_{1s}\right) = \frac{1}{n_{x_1}} \sum_{j=1}^{s} y_j n_{1j}\ ;$$

$$\bar{y}_2 = \frac{1}{n_{x_2}}\left(y_1 n_{21} + y_2 n_{22} + \ldots + y_s n_{2s}\right) = \frac{1}{n_{x_2}} \sum_{j=1}^{s} y_j n_{2j}\ , \ldots;$$

$$\bar{y}_k = \frac{1}{n_{x_2}}\left(y_1 n_{k1} + y_2 n_{k2} + \ldots + y_s n_{ks}\right) = \frac{1}{n_{x_k}} \sum_{j=1}^{s} y_j n_{kj}\ .$$

Графически взаимосвязь двух признаков изображается с помощью *поля корреляции* (рис. 3.1). В прямоугольной системе координат на оси абсцисс откладываются наблюдаемые значения x_i факторного признака X, а на оси ординат — значения y_j результативного признака Y. Затем на плоскости отмечают точки с координатами $\left(x_i, y_j\right)$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, s$, повторяя каждую n_{ij} раз (соответственно частоте данной пары значений), проставляя дополнительные точки вблизи отмеченной.

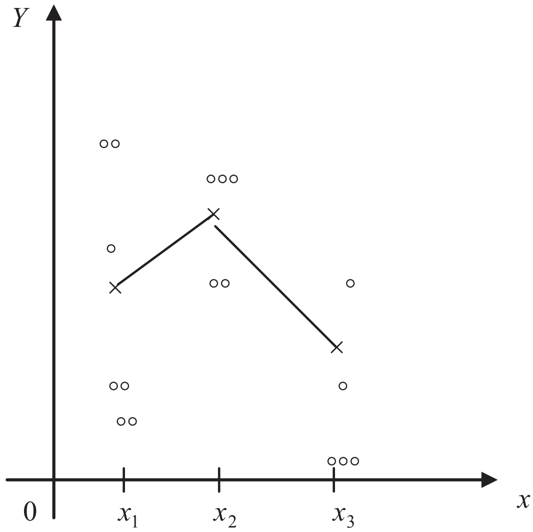


Рис. 3.1. Поле корреляции:
о — точки (x_i, y_i); × — точки $(x_i, \bar{y}_i)$

Если на поле корреляции нанести точки с координатами $(x_i, \bar{y}_i)$, $i = 1, 2, \ldots, k$, и соединить их отрезками, то по полученной ломаной линии можно судить об изменении групповых средних при изменении x и, следовательно, о наличии корреляционной зависимости признака Y от признака X.

Степень разброса значений признака Y относительно условной средней $\bar{y}(x)$ при каждом значении x признака X характеризует тесноту корреляционной зависимости Y от X. Большая концентрация точек на поле корреляции около линии групповых (условных) средних указывает на более сильную корреляционную зависимость между признаками.

3.1. Регрессия. Уравнение регрессии

Рассмотрим две случайные величины (признаки) X, Y с множествами значений $x_1, x_2, \ldots$ и $y_1, y_2, \ldots$ соответственно.

Функция $f(x)$, выражающая корреляционную зависимость признака Y от признака X, то есть описывающая изменение условной средней $\bar{y}(x)$ признака Y при изменении значений признака X, называется *регрессией* Y на X, $\bar{y}(x) = f(x)$.

График функции $\bar{y}(x)$ называется *линией регрессии* случайной величины Y на X.

Теорема (свойство регрессии). *Среди множества всех действительных функций* $f(x)$ *минимум математического ожидания* $M\left(\left[Y - f(x)\right]^2\right)$ *достигается для функции регрессии* $f(x) = \overline{y}(x)$.

Данное свойство используется для аппроксимации зависимости двух случайных величин X и Y приближенной формулой.

Метод наименьших квадратов (МНК)

1. В результате опыта, проведенного в целях установления зависимости двух случайных величин X и Y, получены данные, которые занесены в таблицу:

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

2. Эти же экспериментальные данные представлены графически *диаграммой рассеивания* — набором точек (x_i, y_i) (рис. 3.2).

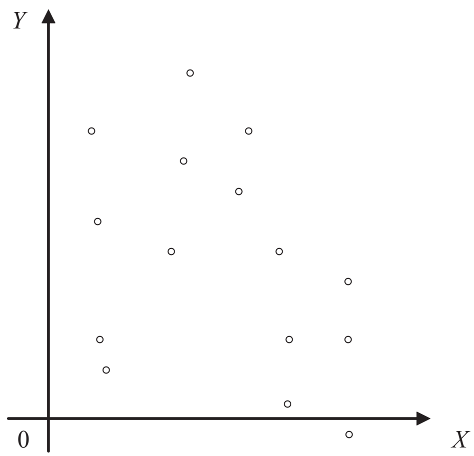


Рис. 3.2. Диаграмма рассеивания

3. Визуальный анализ графических данных является основой предположения о виде функции регрессии

$$f(x) = \overline{y}(x, \alpha_1, \alpha_2, ..., \alpha_l),$$

где α_i — неизвестные параметры, подлежащие определению по данным выборочного наблюдения.

4. МНК-оценки параметров α_i для наилучшего приближения экспериментальных данных находятся следующим образом. Составляют функцию — *сумма квадратов отклонений экспериментальных значений* $\overline{y}_i$ *от соответствующих теоретических значений* $f(x)=\overline{y}(x,\alpha_1,\alpha_2,...,\alpha_l)$:

$$S(\alpha_1,\alpha_2,...,\alpha_l)=\sum_{i=1}^{k}\varepsilon_i^2=\sum_{i=1}^{k}\left(\overline{y}_i-\overline{y}(x_i,\alpha_1,\alpha_2,...,\alpha_l)\right)^2 .$$

Из условия минимума функции $S(\alpha_1,\alpha_2,...,\alpha_l)$ определяют МНК-оценки параметров α_i.

3.2. Линейная регрессия

Линейной регрессией называется функция, которая выражается уравнением прямой

$$\overline{y}(x)=a+bx .$$

Линейная регрессия содержит два параметра. Найдем их оценку по методу наименьших квадратов.

Рассмотрим отклонения экспериментальных значений искомой функции от теоретических (на линейной регрессии):

$$\varepsilon_i=\left(y_i-\overline{y}(x_i,a,b)\right);$$
$$\varepsilon_i=\left(y_i-a-bx_i\right).$$

Составим функцию $S(a,b)$:

$$S(a,b)=\sum_{i=1}^{n}\left(y_i-a-bx_i\right)^2 .$$

Необходимое условие минимума функции $S(a,b)$ приводит к системе уравнений для отыскания оценок a^*,b^*:

$$\begin{cases}\dfrac{\partial S}{\partial a}=0\\[2mm] \dfrac{\partial S}{\partial b}=0,\end{cases}\Rightarrow\begin{cases}-2\sum\limits_{i=1}^{n}\left(y_i-a-bx_i\right)=0,\\ -2\sum\limits_{i=1}^{n}\left(y_i-a-bx_i\right)x_i=0,\end{cases}\Rightarrow\begin{cases}\sum\limits_{i=1}^{n}y_i-an-b\sum\limits_{i=1}^{n}x_i=0,\\ \sum\limits_{i=1}^{n}x_iy_i-a\sum\limits_{i=1}^{n}x_i-b\sum\limits_{i=1}^{n}x_i^2=0.\end{cases}$$

Каждое из уравнений системы домножим на $\frac{1}{n}$ — система приобретает вид

$$\begin{cases} \overline{y} - a - b\overline{x} = 0, \\ \overline{xy} - a\overline{x} - b\overline{x^2} = 0. \end{cases}$$

В итоге получаем формулы для вычисления параметров линейной регрессии

$$b^* = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - \overline{x}^2},$$

$$a^* = \overline{y} - b^*\overline{x}.$$

Покажем, как записать уравнение линейной регрессии в терминах корреляционного анализа, то есть через числовые характеристики выборки двух случайных величин *X* и *Y*.

Перечислим эти характеристики.

- $\mathrm{cov}_{\text{В}}\left(X,Y\right) = \overline{xy} - \overline{x}\ \overline{y}$ — выборочная ковариация;
- $D_{\text{В}}\left(X\right) = \overline{x^2} - \overline{x}^2,\ D_{\text{В}}\left(Y\right) = \overline{y^2} - \overline{y}^2$ — выборочная дисперсия;
- $r_{\text{В}}\left(X,Y\right) = \dfrac{\mathrm{cov}_{\text{В}}\left(X,Y\right)}{\sqrt{D_{\text{В}}(X) \cdot D_{\text{В}}(Y)}}$ — выборочный коэффициент корреляции.

Тогда $b^* = r_{\text{В}}\left(X,Y\right)\sqrt{\dfrac{D_{\text{В}}(Y)}{D_{\text{В}}(X)}}$. Другое обозначение $b^* = \rho_{Y/X} = \dfrac{\overline{xy} - \overline{x}\ \overline{y}}{D_{\text{В}}(X)}$.

Уравнение выборочной линейной регрессии $\overline{y}(x) = a^* + b^*x$ в терминах корреляционного анализа:

$$\overline{y}(x) = \overline{y} + r_{\text{В}}\left(X,Y\right)\sqrt{\frac{D_{\text{В}}(Y)}{D_{\text{В}}(X)}}\left(x - \overline{x}\right),$$

или

$$\overline{y}(x) = \overline{y} + \rho_{Y/X}\left(x - \overline{x}\right),$$

задает прямую, проходящую через точку $\left(\overline{x}, \overline{y}\right)$ с угловым коэффициентом $\rho_{Y/X}$.

Задача 1

В результате наблюдения за уровнем преступности в некотором регионе были получены следующие данные:

Год, x_i	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Количество преступлений, y_i	150	155	130	150	140	125	100	100	90	90

Найти уравнение линейной регрессии.

Решение

Для удобства расчетов примем год 2004 за условную единицу, 2005 год — за двойку, 2006 год — за тройку и так далее. Тогда исходная таблица примет следующий вид:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	150	155	130	150	140	125	100	100	90	90

По заданной выборке вычислим:

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{55}{10} = 5{,}5\,; \quad \overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1230}{10} = 123\,;$$

$$\overline{x^2} = \frac{1}{10}\sum_{i=1}^{10} x_i^2 = \frac{385}{10} = 38{,}5\,; \quad \overline{xy} = \frac{1}{10}\sum_{i=1}^{10} x_i y_i = \frac{6110}{10} = 611\,;$$

$$\rho_{Y/X} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{611 - 5{,}5\cdot 123}{38{,}5 - (5{,}5)^2} = -7{,}94\,.$$

Таким образом, можно записать искомое уравнение регрессии

$$\overline{y}(x) = 123 - 7{,}94\left(x - 5{,}5\right),$$

или

$$\overline{y}(x) = 166{,}67 - 7{,}94x\,.$$

Изобразим графически точки с координатами $\left(x_i, y_i\right)$ и полученную линию регрессии $\overline{y}(x) = 166{,}67 - 7{,}94x$ (рис. 3.3).

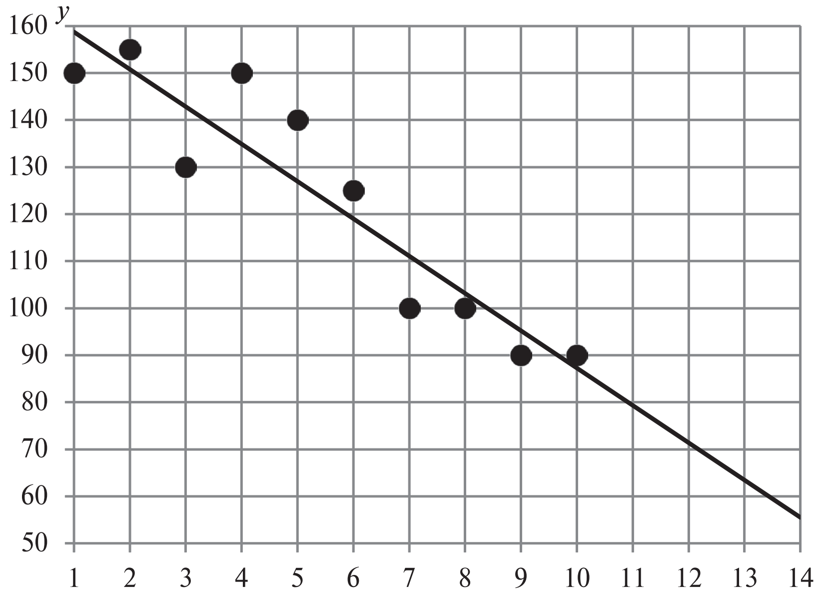


Рис. 3.3. Линия регрессии

Если тенденция сохранится (не произойдет социально-экономических изменений, не будет дефолтов), то мы можем прогнозировать уровень преступности в 2016 году. Для этого нам просто необходимо посчитать значение линейной регрессии *Y* на *X* в этом году, то есть $\bar{y}(13)$. Итак, прогнозируемый уровень преступности в 2016 году будет равен:

$$\bar{y}(x) = 166,67 - 7,94 \cdot 13 = 63 \text{ преступлениям.}$$

3.3. Свойства коэффициента корреляции

Перечислим основные свойства коэффициента:

1. Коэффициент корреляции $r_{\text{В}}(X,Y) = r_{XY}$ принимает значения в следующих пределах: $-1 \le r_{XY} \le 1$.

2. Если случайные величины *X* и *Y* независимы, то есть отсутствует какая-либо зависимость между *X* и *Y* (в том числе корреляционная), то $r_{XY} = 0$.

3. Если $r_{XY} = 0$, то отсутствует линейная корреляционная зависимость *Y* от *X*. Однако может существовать нелинейная зависимость (как корреляционная, так и функциональная).

4. Условие $|r_{XY}|=1$ является необходимым и достаточным для существования линейной функциональной зависимости между Y и X, то есть вида

$$Y = a + bX\ ,\ b \neq 0\ ,$$

где связаны сами случайные величины (а не условная средняя $\bar{y}(x)$ и X в корреляционной зависимости).

3.4. Смысл коэффициента корреляции

Рассмотрим среднюю квадратов отклонений наблюдаемых значений y_i, $i = 1, 2, \ldots, n$ признака Y от линии полученной регрессии $\bar{y}(x) = f(x)$ при всех $X = x_i$, то есть

$$\sigma^2_{\text{регр}} = \frac{1}{n}\sum_{i=1}^{k}\left(y_i - f(x_i)\right)^2.$$

Средней квадратической погрешностью уравнения регрессии Y на X называется величина $\sigma_{\text{регр}} = \sqrt{\sigma^2_{\text{регр}}}$.

В случае линейной регрессии $\bar{y}(x) = \bar{y} + r_{XY}\dfrac{\sigma\ (Y)}{\sigma\ (X)}\left(x - \bar{x}\right)$ средняя квадратическая погрешность

$$\sigma_{\text{регр}} = \sigma_{\text{В}}\left(Y\right)\cdot\sqrt{1 - r^2_{XY}}\ .$$

Таким образом, коэффициент корреляции r_{XY} определяет количественную оценку тесноты линейной корреляционной связи: чем ближе $|r_{XY}|$ к единице ($|r_{XY}| \to 1$), тем теснее линейная корреляционная связь между признаками X и Y ($\sigma_{\text{регр}} \to 0$), то есть значения $\bar{y}_i$ меньше отклоняются от прямой линии регрессии. И наоборот, чем теснее линейная корреляционная зависимость между X и Y, тем ближе $|r_{XY}|$ к единице. Причем условие $|r_{XY}| = 1$ соответствует $\sigma_{\text{регр}} = 0$, то есть все значения y_i признака Y лежат на прямой и, следовательно, Y зависит линейно от X (свойство 3 на с. 70).

Величина r^2_{XY} называется *коэффициентом детерминации* (для линейной связи), который показывает, какую долю дисперсии величины Y можно объяснить зависимостью Y от X (оставшаяся часть дисперсии $\sigma_{\text{В}}{}^2\left(Y\right)$ характеризует степень разброса значений признака Y в зависимости от прочих, кроме X, факторов).

Для получения выводов о количественной оценке тесноты связи применяют *шкалу Чеддока*.

Величина r_{XY}	0,1–0,3	0,3–0,5	0,5–0,7	0,7–0,9	Более 0,9
Характеристика силы связи	Слабая	Умеренная	Заметная	Сильная	Весьма сильная

В случае линейной зависимости в качестве показателя тесноты связи используется $\left|r_{XY}\right|$.

Задача 2

По данным примера 1 (см. с. 69) оценить тесноту линейной корреляционной зависимости количества преступлений Y от года X.

Решение

Вычислим выборочные средние квадратические отклонения $\sigma_В(X)$, $\sigma_В(Y)$:

$$\sigma_В{}^2(X) = \overline{x^2} - \overline{x}^2 = 38{,}5 - (5{,}5)^2 = 8{,}25,\ \sigma_В(X) = \sqrt{\sigma_В{}^2(X)} = \sqrt{8{,}25} = 2{,}872,$$

$$\sigma_В{}^2(Y) = \overline{y^2} - \overline{y}^2 = 15735 - (123)^2 = 606, \sigma_В(Y) = \sqrt{\sigma_В{}^2(Y)} = \sqrt{606} = 24{,}617.$$

Тогда коэффициент корреляции

$$r_{XY} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_В(X)\cdot\sigma_В(Y)} = \frac{611 - 5{,}5\cdot 123}{2{,}872\cdot 24{,}617} = -0{,}926,$$

то есть степень линейной зависимости результативного признака Y от факторного признака X является *очень высокой* (весьма сильная линейная зависимость). Поскольку коэффициент детерминации $r_{XY}^2 = 0{,}858$, то доля дисперсии величины Y, равная 0,858, объясняется зависимостью Y от X. Это означает, что 85,8 % вариации количества преступлений зависит от года.

3.5. Доверительный интервал для линейной регрессии

Поскольку параметры уравнения регрессии определяются по выборочным данным, то их статистические оценки содержат некоторые погрешности (ошибки выборки). Следовательно, величина результативного признака Y окажется с вероятностью γ в определенном интервале относительно значения, вычисленного по уравнению линейной регрессии $\overline{y}(x)$. В случае линейной регрессии доверительные границы, в пределах которых с заданной доверительной вероятностью γ будет находиться теоретическое значение результативного признака Y при значении факторного признака $X = x_0$, определяются следующим образом:

$$\bar{y}_0 - t_\gamma S_{X_0} \le Y \le \bar{y}_0 + t_\gamma S_{X_0},$$

$$S_{X_0} = \frac{\sigma_{\text{регр}}}{\sqrt{n-2}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{\sigma^2(X)}},$$

$$\sigma_{\text{регр}} = \sigma_{\text{В}}(Y) \cdot \sqrt{1 - r_{XY}^2},$$

где $\bar{y}_0$ — значение $\bar{y}(x)$ при $x = x_0$, величина квантили t_γ находится по таблице распределения Стьюдента с доверительной вероятностью γ и числом степеней свободы $k = n - 2$.

Задача 3

Вычислить по данным примера 1 (см. с. 69) с вероятностью 0,95 доверительные границы для количества преступлений, совершенных в 2014 году.

Решение

Имеем $x_0 = 11$. По найденному уравнению регрессии

$$\bar{y}(x) = 166,67 - 7,94x$$

определим $\bar{y}_0$:

$$\bar{y}_0 = 166,67 - 7,94 \cdot 11 = 79,33.$$

Величина квантили t_γ при $\gamma = 0,95$, $k = n - 2 = 10 - 2 = 8$ по таблице распределения Стьюдента (прил. 2) равна $t_\gamma = 2,306$. Поскольку $\sigma(Y) = 24,617$; $\sigma(X) = 2,872$; $r_{XY} = -0,926$; $\bar{x} = 5,5$, то

$$\sigma_{\text{регр}} = 24,617 \cdot \sqrt{1 - (-0,926)^2} = 9,293;$$

$$S_{X_0} = \frac{9,293}{\sqrt{8}} \sqrt{1 + \frac{(11 - 5,5)^2}{8,25}} = 7,097.$$

Таким образом, получим следующий доверительный интервал:

$$79,33 - 2,306 \cdot 7,097 \le Y \le 79,33 + 2,306 \cdot 7,097,$$

или

$$62{,}96 \le Y \le 95{,}69 .$$

Итак, с вероятностью 0,95 можно утверждать, что количество преступлений в 2014 году колеблется от 63 до 96.

3.6. Нелинейная регрессия

В случае значительного отклонения от прямой ломаной линии условных средних применяют нелинейную регрессию $\overline{y}(x) = f(x)$, причем конкретную функцию регрессии $f(x)$ выбирают по виду построенной ломаной.

Средняя квадратическая погрешность $\sigma_{\text{регр}}$ уравнения регрессии Y на X определяется как

$$\sigma_{\text{регр}} = \sqrt{\sigma^2_{\text{регр}}} ,$$

где $\sigma^2_{\text{регр}}$ — среднегрупповая дисперсия значений признака Y относительно линии регрессии $\overline{y}(x) = f(x)$, то есть

$$\sigma^2_{\text{регр}} = \frac{1}{n}\sum_i \sigma^2(x_i) n_{x_i} , \ \sigma^2(x_i) = \frac{1}{n_{x_i}}\sum_j \left(y_j - f(x_i)\right)^2 n_{ij} .$$

Для оценки тесноты связи между признаками X и Y в случае нелинейной зависимости применяют *корреляционное отношение* $\eta_{Y/X}$:

$$\eta_{Y/X} = \sqrt{\frac{\delta^2(Y)}{\sigma^2(Y)}} ,$$

где $\delta^2(Y)$ — *факторная дисперсия* результативного признака Y, характеризующая вариацию Y только в зависимости от воздействия факторного признака X; $\delta^2(Y)$ определяется по формуле межгрупповой дисперсии:

$$\delta^2(Y) = \frac{1}{n}\sum_i \left(\overline{y}_i - \overline{y}\right)^2 n_{x_i} ;$$

$\sigma^2(Y)$ — общая дисперсия результативного признака Y, характеризующая совокупное влияние всех факторов на вариацию Y:

$$\sigma^2(Y) = \frac{1}{n}\sum_j \left(y_j - \bar{y}\right)^2 n_{y_j} \; ;$$

по формуле сложения дисперсий имеем

$$\sigma^2(Y) = \overline{\sigma^2}(Y) + \delta^2(Y) \, ,$$

где $\overline{\sigma^2}(Y)$ — остаточная дисперсия, характеризующая вариацию результативного признака Y от всех прочих, кроме X факторов; $\overline{\sigma^2}(Y)$ определяется по формуле среднегрупповой дисперсии:

$$\overline{\sigma^2}(Y) = \frac{1}{n}\sum_i \sigma_i^2 n_{x_i} \; ,$$

где σ_i^2 — групповая (условная) дисперсия признака Y при $X = x_i, \; i = 1, 2, \ldots, k$:

$$\sigma_i^2 = \frac{1}{n_{x_i}}\sum_j \left(y_j - \bar{y}_i\right)^2 n_{ij} \; .$$

Величина $\eta^2{}_{Y/X}$ называется *коэффициентом детерминации*, который показывает, какая доля общей дисперсии объясняется зависимостью Y от X.

3.7. Свойства корреляционного отношения

1. Корреляционное отношение принимает значения в пределах

$$0 \le \eta_{Y/X} \le 1 \, .$$

2. Равенство $\eta_{Y/X} = 0$ является необходимым и достаточным условием для отсутствия корреляционной зависимости Y от X.
3. Равенство $\eta_{Y/X} = 1$ является необходимым и достаточным условием для функциональной зависимости Y от X.
4. Коэффициент корреляции r_{XY} по модулю не превосходит корреляционного отношения: $|r_{XY}| \le \eta_{Y/X}$, причем $|r_{XY}| = \eta_{Y/X}$ имеет место только в случае линейной регрессии Y на X.

Таким образом, корреляционное отношение $\eta_{Y/X}$ является мерой тесноты линейной корреляционной зависимости *Y* от *X*:

- чем ближе $\eta_{Y/X}$ к единице, тем выше степень корреляционной зависимости *Y* от *X*, которая при $\eta_{Y/X}=1$ переходит в функциональную зависимость; и наоборот, чем выше степень корреляционной зависимости *Y* от *X*, тем ближе $\eta_{Y/X}$ к единице;
- чем ближе $\eta_{Y/X}$ к нулю, тем меньше степень корреляционной зависимости *Y* от *X*, причем при $\eta_{Y/X}=0$ отсутствует корреляционная зависимость *Y* от *X*; и наоборот, слабой корреляционной зависимости *Y* от *X* соответствует близкая к нулю величина $\eta_{Y/X}$.

Корреляционное отношение $\eta_{Y/X}$ является более универсальным показателем тесноты связи по сравнению с коэффициентом корреляции r_{XY}, поскольку используется при любой форме зависимости (коэффициент корреляции применяется только для линейных связей).

Среди различных видов регрессий рассмотрим параболическую и гиперболическую зависимости.

3.8. Параболическая регрессия

Пусть ломаная линия условных средних имеет приблизительно вид параболы. Тогда в качестве регрессии следует выбрать параболическую функцию

$$\bar{y}(x)=a_0+a_1x+a_2x^2,$$

где неизвестные коэффициенты a_0, a_1, a_2 определяются методом наименьших квадратов:

$$S(a_0,a_1,a_2)=\sum_{i=1}\left(\bar{y}_i-a_0-a_1x_i-a_2x_i^{\,2}\right)^2 n_{x_i}\to\min_{a_0,a_1,a_2}.$$

Минимум $S(a_0,a_1,a_2)$ достигается при значениях коэффициентов a_0, a_1, a_2, удовлетворяющих следующей системе уравнений:

$$\begin{cases}a_0+a_1\bar{x}+a_2\overline{x^2}=\bar{y},\\ a_0\bar{x}+a_1\overline{x^2}+a_2\overline{x^3}=\overline{xy},\\ a_0\overline{x^2}+a_1\overline{x^3}+a_2\overline{x^4}=\overline{x^2y},\end{cases}$$

где $\overline{x^k}$ — средняя *k*-х степеней признака *X*, $k=1, 2, 3, 4$:

$$\overline{x^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k n_{x_i} \ ;$$

$\bar{y}$ — средняя значений признака Y; $\overline{xy}$ — средняя произведений значений признаков X и Y; $\overline{x^2 y}$ — средняя произведений квадратов значений признаков X и значений признака Y, то есть

$$\overline{x^2 y} = \frac{1}{n}\sum_i \sum_j x_i^2 y_j n_{ij} \ .$$

3.9. Гиперболическая регрессия

Пусть ломаная линия условных средних имеет приблизительно вид гиперболы. Тогда в качестве регрессии следует выбрать гиперболическую функцию

$$\bar{y}(x) = a_0 + \frac{a_1}{x} \ ,$$

в которой неизвестные коэффициенты a_0, a_1 определяются методом наименьших квадратов. Если предварительно сделать замену переменной $z = \dfrac{1}{x}$, то рассматриваемая функция сводится к линейной функции $\bar{y}(x) = a_0 + a_1 z$, то есть к случаю линейной регрессии. Тогда коэффициенты a_0, a_1 вычисляются по формулам

$$a_0 = \bar{y} - a_1 \bar{z} \ , \ a_1 = r_{ZY} \cdot \frac{\sigma(Y)}{\sigma(Z)} \ ,$$

где $\bar{y}$, $\bar{z}$ — средние значений признаков Y и $Z = \dfrac{1}{X}$ соответственно; $\sigma(Y)$, $\sigma(Z)$ — средние квадратические отклонений этих признаков; r_{ZY} — коэффициент корреляции:

$$r_{ZY} = \frac{\overline{zy} - \bar{z}\ \bar{y}}{\sigma(Z)\cdot\sigma(Y)} \ .$$

Таким образом, уравнение гиперболической регрессии Y на X имеет вид

$$\bar{y}(x) = \bar{y} + r_{ZY} \frac{\sigma(Y)}{\sigma(Z)} \left(\frac{1}{x} - \bar{z} \right).$$

Задача 4

Для установления зависимости доли легальных предприятий (Y) от величины процентной ставки налога на прибыль (X) проведено выборочное наблюдение, в результате которого получены следующие расчетные данные (данные условные).

Налог на прибыль X, %	1	1,5	2	3	4	6	8	9	10
Доля легальных предприятий Y, %	91	70	53	41	28	20	16	15	15

Найти уравнение регрессии Y на X, выбрав функцию регрессии по виду ломаной наблюдаемых значений Y, изобразить ее графически.

Решение

Нанесем на координатную плоскость XY точки: по оси X будем откладывать налог, по оси Y — долю легальных предприятий (рис. 3.4).

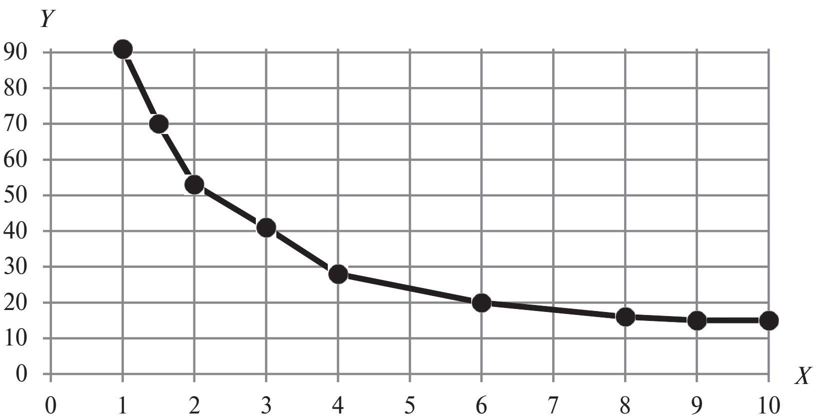


Рис. 3.4. Ломаная наблюдаемых значений признака Y

По виду ломаной наблюдаемых значений Y можно предположить наличие зависимости Y от X, причем допустимо выбрать функцию регрессии гиперболического типа:

$$\bar{y}(x) = a_0 + \frac{a_1}{x}.$$

Сделаем замену переменной $z = \dfrac{1}{x}$, тогда рассматриваемая функция сводится к линейной функции $\bar{y}(x) = a_0 + a_1 z$, то есть к случаю линейной регрессии. Найдем неизвестные коэффициенты a_0, a_1. По данным задачи вычислим средние величины:

$$\bar{y} = \frac{1}{n}\sum_i y_i = \frac{1}{9}(91+70+53+41+28+20+16+15+15) = 38,778\,;$$

$$\bar{z} = \frac{1}{n}\sum_i z_i = \frac{1}{9}\left(1+\frac{2}{3}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\frac{1}{9}+\frac{1}{10}\right) = 0,361\,;$$

$$\overline{zy} = \frac{1}{n}\sum_i z_i y_i = \frac{1}{9}\left(1\cdot 91+\frac{2}{3}\cdot 70+\frac{1}{2}\cdot 53+\frac{1}{3}\cdot 41+\frac{1}{4}\cdot 28\right)+$$

$$+\frac{1}{9}\left(\frac{1}{6}\cdot 20+\frac{1}{8}\cdot 16+\frac{1}{9}\cdot 15+\frac{1}{10}\cdot 15\right) = 21,481\,;$$

$$\overline{z^2} = \frac{1}{n}\sum_i z_i^{\,2} = \frac{1}{9}\left(1+\frac{4}{9}+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\frac{1}{36}+\frac{1}{64}+\frac{1}{81}+\frac{1}{100}\right) = 0,215\,.$$

Рассчитаем выборочную дисперсию $\sigma^2(Z)$:

$$\sigma^2(Z) = \overline{z^2} - \bar{z}^2 = 0,215-(0,361)^2 = 0,084\,,$$

тогда выборочное среднее квадратическое:

$$\sigma(Z) = \sqrt{\sigma^2(Z)} = \sqrt{0,084} = 0,290\,.$$

Подставим найденные величины в формулы коэффициентов a_1 и a_0:

$$a_1 = \frac{\overline{zy}-\bar{z}\,\bar{y}}{\sigma^2(Z)} = \frac{21,481-0,361\cdot 38,778}{0,084} = 89,073\,,$$

$$a_0 = \bar{y} - a_1\bar{z} = 38,778-89,073\cdot 0,361 = 6,622\,.$$

Таким образом, уравнение гиперболической регрессии *Y* на *X* имеет вид

$$\bar{y}(x) = 6,745+\frac{89,073}{x}\,.$$

Изобразим графически точки с координатами (x_i, y_i) и полученное уравнение регрессии $\overline{y}(x) = 6{,}745 + \dfrac{89{,}073}{x}$ (рис. 3.5).

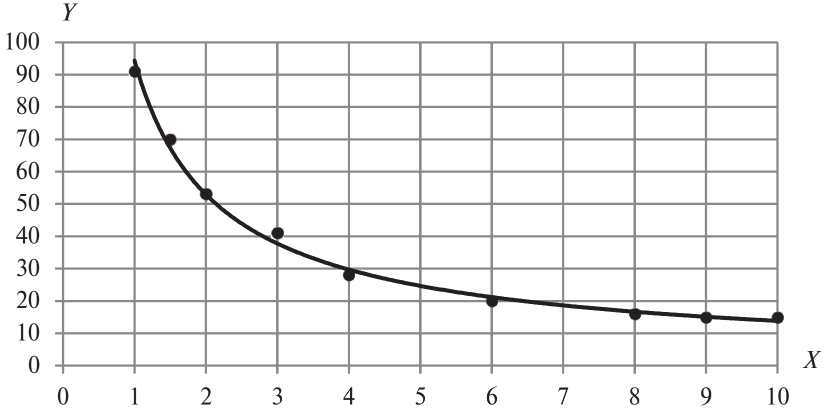


Рис. 3.5. Гиперболическая регрессия Y на X

Если тенденция сохранится, то мы можем прогнозировать долю легальных предприятий при увеличении налоговой ставки на прибыль до 16 %. Для этого нам просто необходимо посчитать значение гиперболической регрессии Y на X при такой налоговой ставке, то есть $\overline{y}(16)$. Итак, рассчитаем прогнозируемую долю легальных предприятий при значении налоговой ставки на прибыль 16 %:

$$\overline{y}(x) = 6{,}745 + \frac{89{,}073}{16} = 12\ \%.$$

3.10. Построение линейной регрессии по несгруппированным данным

Опишем методику проведения лабораторной работы по решению задачи корреляционно-регрессионного анализа.

Цель работы:

- овладение способами построения моделей линейной корреляции для несгруппированных данных;
- вычисление основных числовых характеристик выборки;
- построение уравнений регрессии, диаграммы рассеивания.

Задача 5

Для установления зависимости между двумя признаками *X* и *Y* произведено статистическое наблюдение, результаты которого приведены в таблице:

X	10	8	13	9	11	14	6	4	12	7	5
Y	7,46	6,76	12,75	7,11	7,81	8,84	6,10	5,39	8,14	6,42	5,73

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что *X* и *Y* связаны линейной зависимостью, написать *уравнения линейной регрессии Y* на *X* и *X* на *Y*, вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Порядок выполнения работы

Запуск Microsoft Excel

Запустить **Microsoft Excel**, в новом документе удалить листы 2 и 3, сохранить файл в своей папке под именем «Задача 3». Далее нужно периодически сохранять документ.

Переименовать *Лист1* в *Расчетная таблица*. Расчетная таблица должна иметь следующий вид (рис. 3.6):

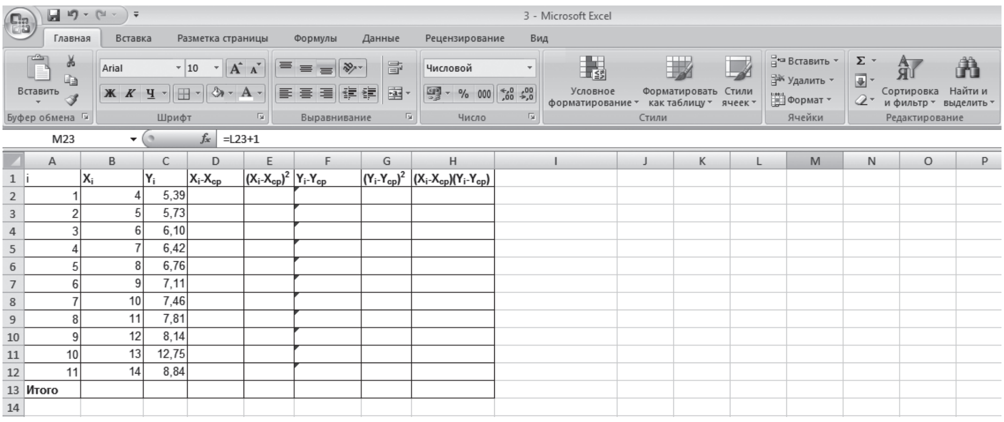


Рис. 3.6. Расчетная таблица

Заполнение расчетной таблицы

Набрать заголовки в первой строке. Чтобы сделать текст нижним или верхним индексом, можно действовать так: сначала набрать в ячейке B1 просто *Xi*, затем в строке формул выделить *i*, выбрать на вкладке *Главная Шрифт*, там, в разделе *Видоизменение*, установить галочку напротив надписи «подстрочный». Аналогично делаются индексы в остальных ячейках. Сделать заголовки полужирным, выравнивание по центру.

Ввести данные, известные из условия задачи: индекс i, значения признака X — X_i и признака Y — Y_i .

В строке *Итого* в столбцах B и C вычислить итоговые суммы. Для этого установить курсор мыши на ячейке B13 и нажать кнопку «*Автосумма*» на панели инструментов вкладки *Формула*. Ячейки C13 заполнить с помощью маркера заполнения.

Вычисление числовых характеристик выборки

Строку 15 отвести под вычисление объема выборки n. В ячейке A15 набрать текст n=. Объем выборки вычислить в ячейке B15, используя функцию *СЧЁТ*.

Итак, перейти в ячейку B15, на панели инструментов вкладки *Формула* нажать кнопку «*Вставить функцию* f_x». В открывшемся окне *Мастера функций* выбрать в списке категорий *Статистические*, в списке функций, используя полосу прокрутки, выбрать *СЧЁТ*, щелкнуть *OK,* далее в окне *Аргументы функции* (рис. 3.7) в поле *Значение 1* ввести диапазон адресов ячеек A2: A12, нажать *OK*.

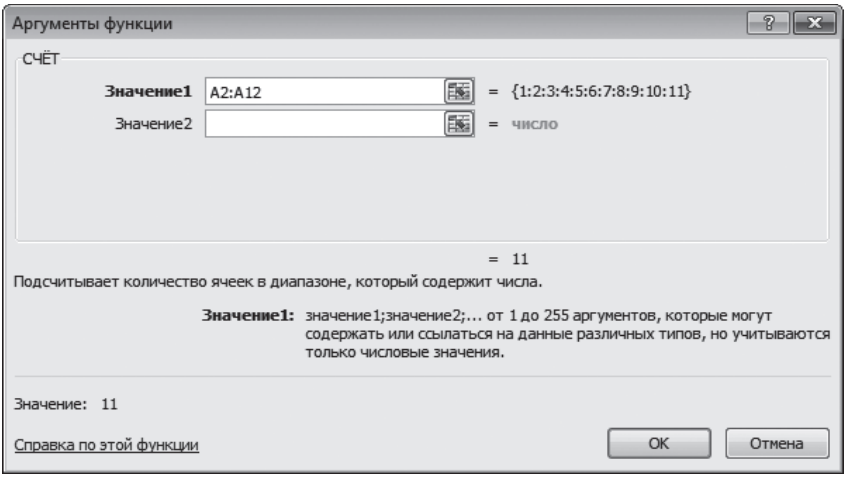


Рис. 3.7. Окно *Аргументы функций*

В строке 16 вычислить средние значения признаков X и Y.

В ячейке A16 ввести $X_{ср}$ =. Выборочное среднее признака X вычисляется по формуле

$$X_{ср} = \frac{1}{n} \sum_i X_i \; .$$

Суммарные значения X_i были вычислены в ячейке B13, объем выборки n — B15. Таким образом, в ячейку B16 ввести следующую формулу:

$$= B13 / B15 .$$

В ячейке D16 ввести $Y_{\text{ср}} =$. Выборочное среднее признака Y вычисляется по формуле

$$Y_{\text{ср}} = \frac{1}{n}\sum_i Y_i .$$

Суммарные значения X_i были вычислены в ячейке C13, объем выборки n — B15. Таким образом, в ячейку E16 ввести следующую формулу:

$$= C13 / B15 .$$

Далее заполните столбцы D и E по формулам

$$X_i - X_{\text{ср}} ,$$

$$\left(X_i - X_{\text{ср}}\right)^2$$

соответственно. Для этого в ячейку D2 введите следующую расчетную формулу:

$$= B2 - \$B\$16 ,$$

в ячейку E2 — формулу

$$= D2 \hat{} 2 .$$

Остальные ячейки столбцов D и E заполнить с помощью маркера заполнения.

Затем заполнить столбцы F и G по формулам

$$Y_i - Y_{\text{ср}} ,$$

$$\left(Y_i - Y_{\text{ср}}\right)^2$$

соответственно. Для этого в ячейку F2 ввести следующую расчетную формулу:

$$= C2 - \$E\$16 ,$$

в ячейку G2 — формулу

$$= F2 \hat{} 2 .$$

Остальные ячейки столбцов F и G заполнить с помощью маркера заполнения.

Теперь заполнить столбец H по формуле

$$\left(X_i - X_{\text{ср}}\right)\left(Y_i - Y_{\text{ср}}\right).$$

В ячейку H2 ввести следующую расчетную формулу:

$$= D2 * F2 .$$

Остальные ячейки столбца H заполнить с помощью маркера заполнения.

В строке *Итого* в столбцах D, E, F, G, H вычислить итоговые суммы.

Перейти к вычислению *выборочных дисперсий* признаков *X* и *Y*: их необходимо рассчитать в строке 17.

В ячейке A17 ввести $D_X =$. Выборочная дисперсия вычисляется по формуле

$$D_X = \frac{1}{n}\sum_i \left(X_i - X_{\text{ср}}\right)^2 .$$

Суммарные значения $\left(X_i - X_{\text{ср}}\right)^2$ были вычислены в ячейке E13, объем выборки *n* — B15. Таким образом, в ячейку B17 ввести следующую формулу

$$= E13 / B15 .$$

В ячейке D17 ввести $D_Y =$. Выборочная дисперсия вычисляется по формуле

$$D_Y = \frac{1}{n}\sum_i \left(Y_i - Y_{\text{ср}}\right)^2 .$$

Суммарные значения $\left(Y_i - Y_{\text{ср}}\right)^2$ были вычислены в ячейке G13, объем выборки *n* — B15. Таким образом, в ячейку E17 введите следующую формулу:

$$= G13 / B15 .$$

Теперь вычислим выборочную ковариацию K_{XY}. В ячейке A18 ввести $K_{XY} =$. Выборочная ковариация вычисляется по формуле

$$K_{XY} = \frac{1}{n}\sum_i \left(X_i - X_{\text{ср}}\right)\left(Y_i - Y_{\text{ср}}\right).$$

Суммарные значения произведений $\left(X_i - X_{\text{ср}}\right)$ на $\left(Y_i - Y_{\text{ср}}\right)$ были вычислены в ячейке H13, объем выборки n — B15. Таким образом, в ячейку B18 ввести следующую формулу:

$$= H13 / B15.$$

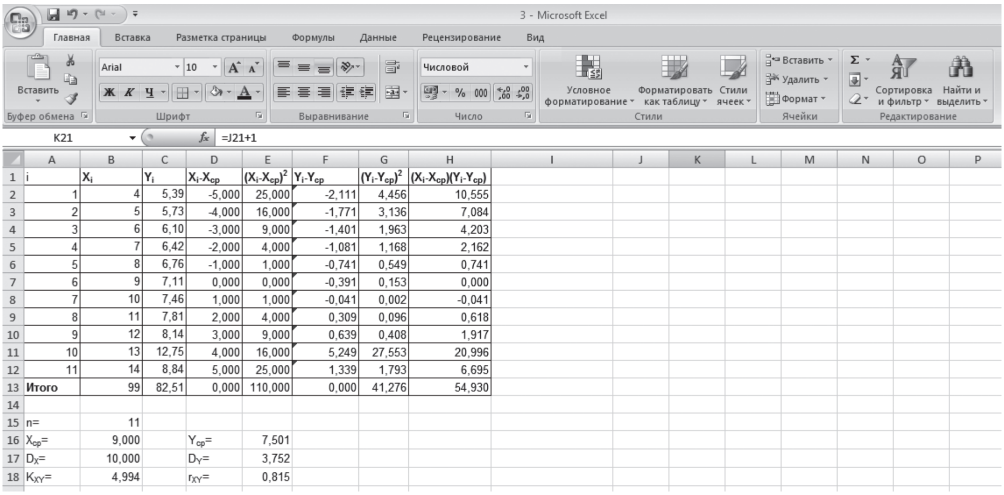
Найти выборочный коэффициент корреляции r_{XY}. В ячейке D18 ввести $r_{XY} =$. Выборочный коэффициент корреляции вычислить по формуле

$$r_{XY} = \frac{K_{XY}}{\sqrt{D_X D_Y}}.$$

Таким образом, в ячейку E18 ввести следующую формулу:

$$= B18 / \text{КОРЕНЬ}(B17 * E17).$$

На данном этапе должны получиться результаты вычислений, как на рис. 3.8.



	A	B	C	D	E	F	G	H
1	i	X_i	Y_i	X_i-X_{cp}	$(X_i$-$X_{cp})^2$	Y_i-Y_{cp}	$(Y_i$-$Y_{cp})^2$	$(X_i$-$X_{cp})(Y_i$-$Y_{cp})$
2	1	4	5,39	-5,000	25,000	-2,111	4,456	10,555
3	2	5	5,73	-4,000	16,000	-1,771	3,136	7,084
4	3	6	6,10	-3,000	9,000	-1,401	1,963	4,203
5	4	7	6,42	-2,000	4,000	-1,081	1,168	2,162
6	5	8	6,76	-1,000	1,000	-0,741	0,549	0,741
7	6	9	7,11	0,000	0,000	-0,391	0,153	0,000
8	7	10	7,46	1,000	1,000	-0,041	0,002	-0,041
9	8	11	7,81	2,000	4,000	0,309	0,096	0,618
10	9	12	8,14	3,000	9,000	0,639	0,408	1,917
11	10	13	12,75	4,000	16,000	5,249	27,553	20,996
12	11	14	8,84	5,000	25,000	1,339	1,793	6,695
13	**Итого**	99	82,51	0,000	110,000	0,000	41,276	54,930
14								
15	n=	11						
16	X_{cp}=	9,000		Y_{cp}=	7,501			
17	D_X=	10,000		D_Y=	3,752			
18	K_{XY}=	4,994		r_{XY}=	0,815			

Рис. 3.8. Результаты вычислений

Составление уравнения линейной регрессии

Выборочная линейная регрессия Y на X по выборке $\left(X_i, Y_i\right)$ определяется уравнением

$$\overline{y}(x) = \beta_0 + \beta_1 x = Y_{\text{ср}} + r_{XY} \sqrt{\frac{D_Y}{D_X}} \left(x - X_{\text{ср}}\right),$$

где выборочные коэффициенты регрессии:

$$\beta_1 = \frac{K_{XY}}{D_X},$$

$$\beta_0 = Y_{\text{ср}} - \beta_1 X_{\text{ср}}.$$

Приступить к вычислению выборочных коэффициентов регрессии.

В ячейках A20 и A21 ввести $\beta_1 =$, $\beta_0 =$ соответственно.

В ячейку B20 ввести следующую формулу:

$$= B18 / B17.$$

В ячейку B21 ввести следующую формулу:

$$= E16 - B20 * B16.$$

Аналогично определяется уравнение выборочной линейной регрессии X на Y по выборке (X_i, Y_i):

$$\bar{x}(y) = \beta_0' + \beta_1' y = X_{\text{ср}} + r_{XY}\sqrt{\frac{D_X}{D_Y}}\left(y - Y_{\text{ср}}\right),$$

где выборочные коэффициенты регрессии

$$\beta_1' = \frac{K_{XY}}{D_Y},$$

$$\beta_0' = X_{\text{ср}} - \beta_1' Y_{\text{ср}}.$$

Вычислим выборочные коэффициенты регрессии X на Y.

В ячейках A22 и A23 ввести $\beta_1' =$, $\beta_0' =$ соответственно.

В ячейку B22 ввести следующую формулу:

$$= B18 / E17.$$

В ячейку B23 ввести следующую формулу:

$$= B16 - B22 * E16.$$

В итоге должны получиться значения коэффициентов регрессии, как на рис. 3.9.

β_1	0,49936
β_0	3,00664
β_1'	1,33078
β_0'	-0,98207

Рис. 3.9. Коэффициенты регрессии

Таким образом, уравнение выборочной линейной регрессии Y на X по выборке (X_i, Y_i) имеет вид

$$\bar{y}(x) = 3,00664 + 0,49936x\ ,$$

уравнение выборочной линейной регрессии X на Y по выборке (X_i, Y_i) имеет вид

$$\bar{x}(y) = -0,98207 + 1,33078y\ .$$

Вычислить среднюю квадратическую погрешность уравнения линейной регрессии Y на X по формуле

$$\sigma_{\text{регр}} = \sqrt{D_Y}\sqrt{1 - r_{XY}^2}\ .$$

В ячейке A19 сделать подпись $SI_{\text{регр}} =$. В ячейку B19 ввести следующую формулу:

$$= \text{КОРЕНЬ}\left(E17 * (1 - E18)\,\hat{}\,2\right).$$

Теперь необходимо вычислить среднюю квадратическую погрешность уравнения линейной регрессии X на Y по формуле

$$\sigma_{\text{регр}} = \sqrt{D_X}\sqrt{1 - r_{XY}^2}\ .$$

В ячейку D19 ввести $SI'_{\text{регр}}$, а в ячейку E19 — следующую формулу:

$$= \text{КОРЕНЬ}\left(B17 * (1 - E18)\,\hat{}\,2\right).$$

Построение диаграмм

Для построения линейной регрессии Y на X и X на Y заполнить ячейки I1, J1, K1 расчетной таблицы следующим образом:

- в ячейку I1 введите заголовок *t*;
- в ячейку J1 — $\bar{y}(t)$;
- в ячейку K1 — $\bar{x}(t)$.

В ячейке I2 набрать 3, в ячейке I3 — 4, далее выделить ячейки I2 и I3, подвести указатель мыши к маркеру заполнения и, удерживая левую кнопку мыши, потащить вниз до ячейки I19.

Теперь необходимо табулировать значение функции выборочной регрессии *Y* на *X*:

$$\bar{y}(x) = \beta_0 + \beta_1 x .$$

Перейти в ячейку J2 и ввести в нее формулу

$$= \$B21 + \$B20 * I2 .$$

Далее заполнить значение функции выборочной регрессии *X* на *Y*:

$$\bar{x}(y) = \beta_0' + \beta_1' y .$$

Перейти в ячейку K2 и ввести в нее формулу

$$= \$B23 + \$B22 * I2 .$$

Отсортировать столбец значений признака X_i по возрастанию. Для этого выделить диапазон ячеек B2 : B12, перейти на вкладку *Данные* и щелкнуть по кнопке «*Сортировка*».

Перейти к построению графиков функций регрессий и диаграммы рассеивания.

На вкладке *Вставка* в строке меню *Диаграммы* выбрать вид диаграммы *Точечная,* затем подвид диаграммы *Точечная с прямыми отрезками* (справа во втором ряду, рис. 3.10).

В результате на рабочем листе появится окно с пустой диаграммой. Для отображения на ней данных выполнить следующие действия:

1. В контекстном меню диаграммы выполнить пункт *Выбрать данные*.

2. В открывшемся окне в поле *Диапазон данных для диаграммы* указать ячейки, в которых находятся значения $\bar{y}(t)$. В разделе *Элементы легенды* нажать *Изменить* и указать в поле *Значения X* ячейки, в которых находятся значения *t*. Нажать *OK*.

3. Щелкнуть по кнопке *Добавить* в разделе *Элементы легенды* и указать в поле *Значения X* ячейки, в которых находятся значения *t*, в поле *Значения Y* ячейки, где находятся ***значения*** $\bar{x}(t)$. Нажать *OK*.

4. Еще раз нажать *Добавить* в разделе *Элементы легенды* и указать в поле *Значения X* ячейки, в которых находятся значения X_i, в поле *Значения Y* ячейки, в которых находятся значения Y_i. Нажать *OK*.

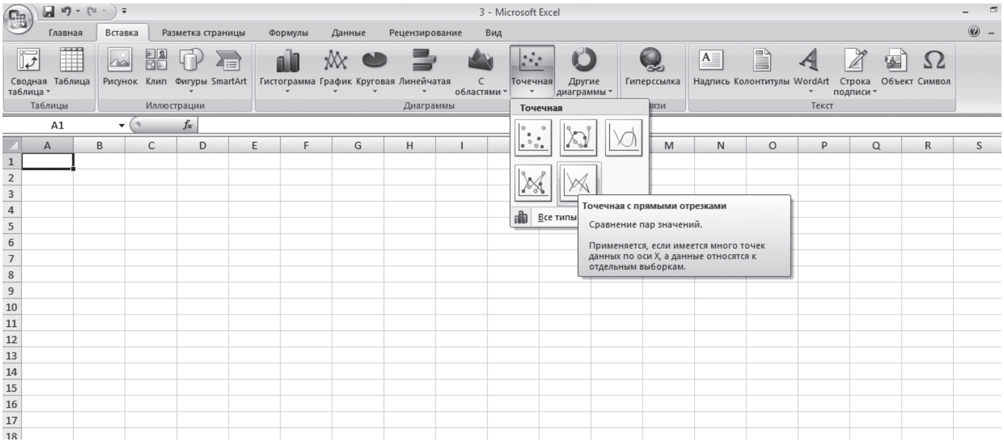


Рис. 3.10. Точечная диаграмма

В результате будет построено три графика в одном окне диаграммы (рис. 3.11).

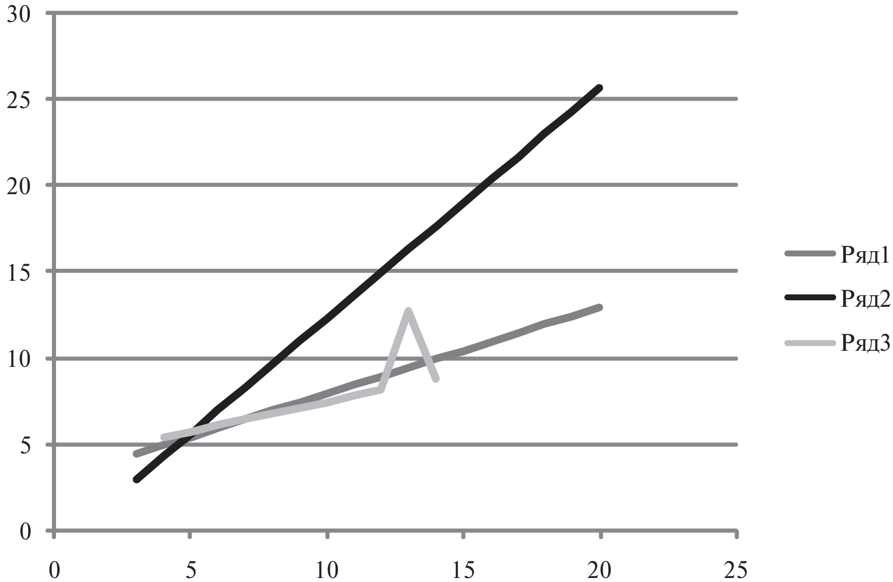


Рис. 3.11. Три ряда диаграммы

Расположить диаграмму на отдельном листе с именем *Диаграмма рассеивания*. Для этого щелкнуть диаграмму, вызвать контекстное меню и выполнить пункт *Переместить диаграмму*. В появившемся окне в разделе *Разместить диаграмму* установить переключатель в положение *На отдельном листе* и ввести имя нового листа диаграммы в поле этого переключателя. Нажать *OK*.

Щелкнуть по диаграмме, в командной строке *Работа с диаграммами* на вкладке *Макет* в группе *Фон* нажать на кнопку *Область построения*, выбрать в выпадающем меню вариант *Нет* (*удалить заливку области построения*).

В группе *Оси* кнопки *Оси* и *Сетка* позволяют настроить отображение осей и линий сетки для диаграммы. Опции *Основная горизонтальная ось* и *Основная вертикальная ось* выбираются по умолчанию. Создайть и горизонтальные, и вертикальные линии сетки для основных делений — опция *Основные линии сетки*.

Вызвать контекстное меню (щелкнуть правой кнопкой мыши) на линии регрессии X на Y, выбрать команду *Формат ряда данных*, пункт *Тип линии*, в подпункте *Тип штриха* установить *Штрих* (рис. 3.12).

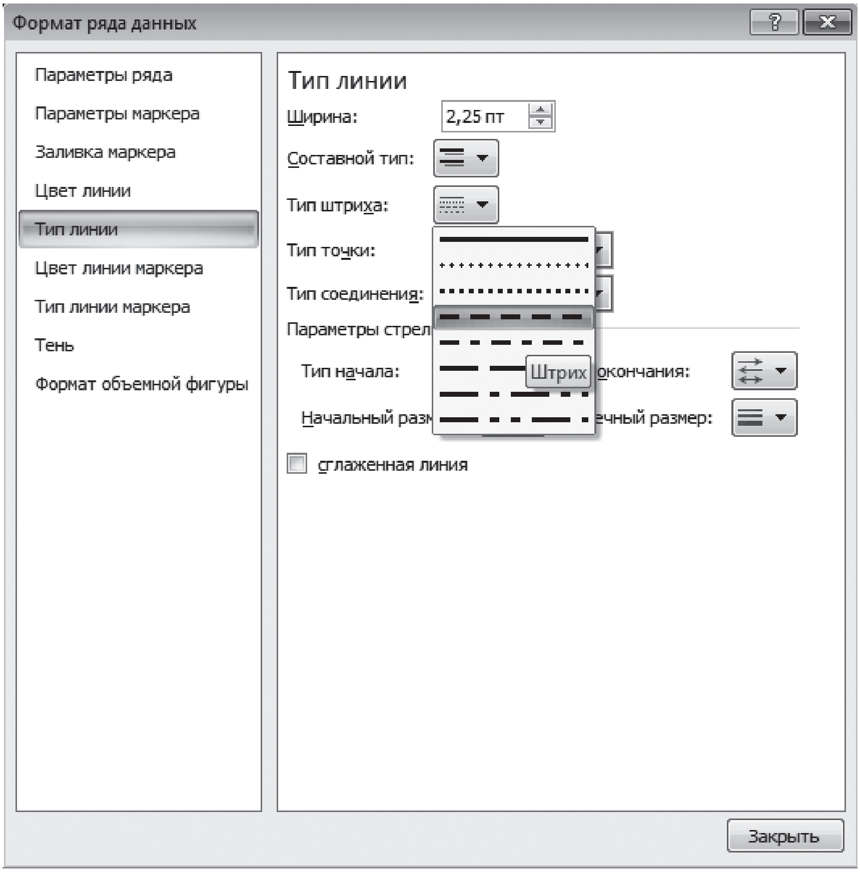


Рис. 3.12. Формат ряда данных

Вызвать контекстное меню на линии рассеивания (X_i, Y_i), выбрать команду *Изменить тип диаграммы для ряда*, в открывшемся окне *Изменение типа диаграммы* выбрать *Точечная с маркерами* (рис. 3.13).

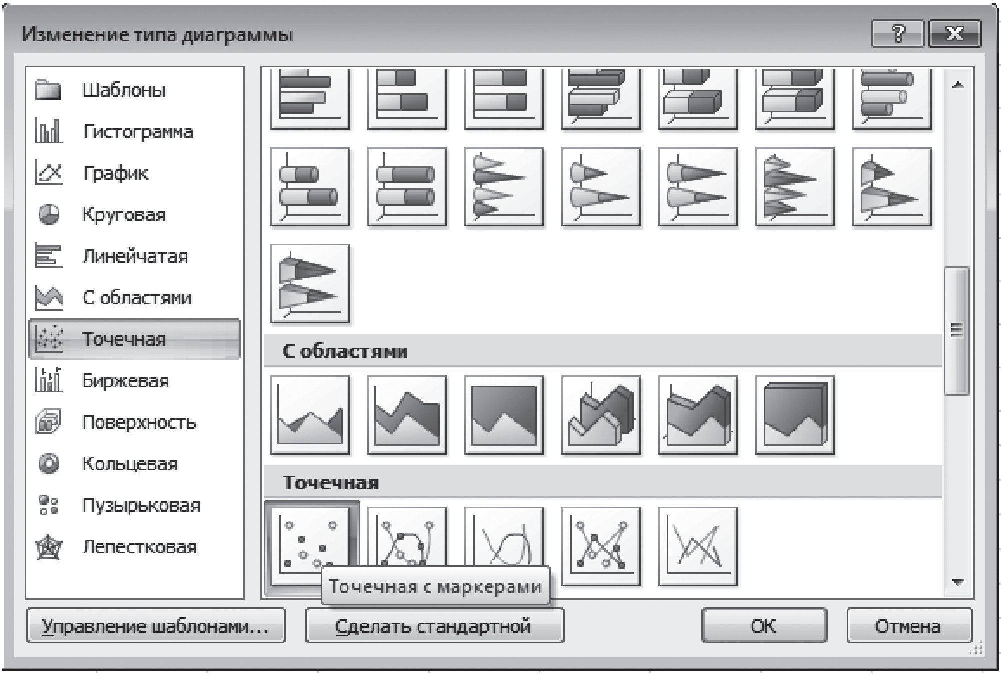


Рис. 3.13. Изменение типа диаграммы

Изменить тип маркера построенной точечной диаграммы на «кружок». Для этого снова вызвать контекстное меню построенных маркеров, выбрать команду *Формат ряда данных*, в пункте *Параметры маркера* установить тип *кружок*.

Добавить значения данных на диаграмме рассеивания. Щелкнуть правой кнопкой мыши на диаграмме рассеивания, выбрать пункт *Добавить подписи данных*.

Для настройки параметров горизонтальной и вертикальной осей (форматирование этих элементов диаграммы) следует перейти к окну *Формат оси* с помощью контекстного меню. Для горизонтальной и вертикальной оси выбрать на вкладке *Параметры оси* минимальное значение *3*, максимальное значение *20*, цена основных делений *1*.

Щелкнуть по диаграмме, в командной строке *Работа с диаграммами* на вкладке *Макет* в группе команд *Подписи* нажать кнопку *Названия осей*, выбрать *Название основной горизонтальной оси*, затем вариант *Название под осью*. Ввести вместо *Названия оси* обозначение X_i, затем перетащить его (методом «буксировки») в положение справа от оси. Выполнить аналогичные действия для названия основной вертикальной оси, введя обозначения для нее Y_i и перетащив его в положение над осью. Щелкнув кнопку *Легенда*, выбрать вариант *Нет*.

На рис. 3.14 приведены графики, которые должны получиться.

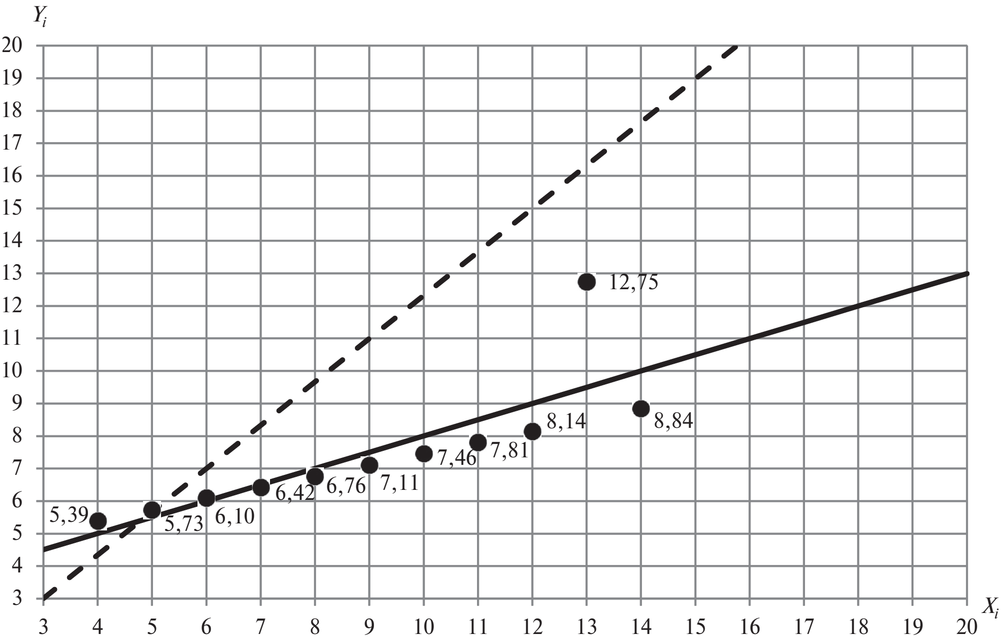


Рис. 3.14. Линии регрессии:
сплошная прямая — регрессия Y на X; штриховая — регрессия X на Y;
точки — диаграмма рассеивания

Индивидуальное задание

Создайть лист *Расчетная таблица №*. Решить свой вариант задания.

Результаты работы

В результате выполненной работы студент должен продемонстрировать преподавателю готовый файл *Задача 3.xlsx*, содержащий 5 листов:

- лист 1 (*Титульный лист*) — титульный лист к работе, на котором указаны название работы, номер варианта, Ф. И.О. студента, номер группы, дата выполнения работы;
- лист 2 (*Расчетная таблица*) — таблица исходных данных и решения задачи;
- лист 3 (*Диаграмма рассеивания*) — диаграмма рассеивания, линии регрессии задачи;
- лист 4 (*Расчетная таблица №...*) — таблица исходных данных и решения задания своего варианта;
- лист 5 (*Диаграмма рассеивания №...*) — диаграмма рассеивания, линии регрессии задания своего варианта.

Варианты индивидуального задания

Вариант 1

Для установления зависимости между двумя признаками X и Y произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	10	8	13	9	11	14	6	4	12	7	5
Y	8,04	6,95	7,58	8,82	8,33	9,96	7,24	4,26	10,83	4,81	5,68

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что X и Y связаны линейной зависимостью, написать уравнения линейной регрессии Y на X и X на Y. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Вариант 2

Для установления зависимости между двумя признаками X и Y произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	10	8	13	9	11	14	6	4	12	7	5
Y	9,14	8,17	8,74	8,77	9,26	8,10	6,11	3,11	9,13	7,22	4,73

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что X и Y связаны линейной зависимостью, написать уравнения линейной регрессии Y на X и X на Y. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Вариант 3

Для установления зависимости между двумя признаками X и Y произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	8	8	8	8	8	8	19	8	8	8	8
Y	5,21	6,05	6,75	6,245	8,465	5,97	12,49	8,79	8,26	6,37	7,89

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что X и Y связаны линейной зависимостью, написать уравнения линейной регрессии Y на X и X на Y. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Вариант 4

Для установления зависимости между двумя признаками X и Y произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	1	2	3	4	5	6	7	8	9
Y	61,8	56,1	50,8	47,6	42,3	39,5	35,7	29,4	25,1

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что X и Y связаны линейной зависимостью, написать уравнения линейной регрессии Y на X и X на Y. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Вариант 5

Для установления зависимости между двумя признаками X и Y произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	50	60	75	80	100	120	150	300	500	1000
Y	10,6	10,4	10,3	10,1	10,0	9,7	9,5	7,5	6,0	1,1

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что X и Y связаны линейной зависимостью, написать уравнения линейной регрессии Y на X и X на Y. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Вариант 6

Для установления зависимости между двумя признаками X и Y произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	1	2	3	4	5	6	7	8	9	10
Y	20	25	31	31	40	56	52	60	60	70

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что *X* и *Y* связаны линейной зависимостью, написать уравнения линейной регрессии *Y* на *X* и *X* на *Y*. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Вариант 7

Для установления зависимости между двумя признаками *X* и *Y* произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	50	60	75	80	100	120	150	300	500	1000
Y	21,2	20,8	20,6	20,2	20,0	19,4	19,0	15,0	12,0	2,2

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что *X* и *Y* связаны линейной зависимостью, написать уравнения линейной регрессии Y на *X* и *X* на *Y*. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Вариант 8

Для установления зависимости между двумя признаками *X* и *Y* произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	406	660	914	1168	1422	1676	1930	2184	2438
Y	518,5	813,5	1108,5	1403,5	1698,05	1993,5	2288,5	2583,8	2878,5

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что *X* и *Y* связаны линейной зависимостью, написать уравнения линейной регрессии *Y* на *X* и *X* на *Y*. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

Вариант 9

Для установления зависимости между двумя признаками *X* и *Y* произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	50	49	48	51	52	53	54	57	59	60	61	55
Y	10	8	10	9	10	12	13	15	16	18	20	17

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что X и Y связаны линейной зависимостью, написать уравнения линейной регрессии Y на X и X на Y. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, повести прямые линейной регрессии.

Вариант 10

Для установления зависимости между двумя признаками X и Y произведено статистическое наблюдение, результаты которого приведены в следующей таблице:

X	39	42	53	70	73,5	75	90	98	110	115
Y	1,3	1,3	0,8	2,2	1,8	2	2,2	1,8	2,8	2,1

Найти числовые характеристики выборки: среднее, дисперсию, ковариацию, коэффициент корреляции.

Предполагая, что X и Y связаны линейной зависимостью, написать уравнения линейной регрессии Y на X и X на Y. Вычислить среднюю квадратическую погрешность полученных уравнений регрессии.

Построить диаграммы рассеяния, провести прямые линейной регрессии.

4. Статистическая проверка гипотез

В процессе анализа данных статистического наблюдения выдвигаются различные предположения вероятностного характера относительно генеральной совокупности. Например, в качестве возможного выбирается конкретный вид распределения изучаемого признака, указывается его аналитическая форма представления; исследуемая причинно-следственная связь изображается с помощью определенной модели, включающей в себя лишь часть факторных признаков (объявленных в рамках этой модели наиболее существенными); аналитически описывается предполагаемая тенденция развития процесса т. д. Подобные суждения, теоретически выражающие статистические зависимости и закономерности, называются *статистическими гипотезами*. Их проверка заключается в оценке существенности расхождений между теоретическими и эмпирическими (то есть вычисленными по данным наблюдения) показателями с помощью специальных статистических методов, называемых *критериями согласия*. Надежность оценки характеризуется некоторой вероятностью.

Обозначим через H_0 предположение, подлежащее статистической проверке. Тогда H_0 называют *основной* (*нулевой*) *гипотезой*. Пусть H_1 — *альтернативная гипотеза*, конкурирующая с H_0 и принимаемая в случае, если в результате статистической проверки гипотеза H_0 отвергается. Решение о справедливости основной гипотезы H_0 или альтернативной H_1 принимается по выборочным данным (полученным в результате случайного отбора) и, следовательно, может быть ошибочным (с некоторой вероятностью):

1) отвергается правильная гипотеза H_0 (ошибка *первого рода*);
2) принимается неправильная гипотеза H_0 (ошибка *второго рода*).

Уровень значимости α — вероятность ошибки первого рода. Величина α задается заранее и определяет надежность принятого решения: с вероятностью $\gamma = 1 - \alpha$ верная гипотеза H_0 будет принята. Вероятность ошибки второго рода обозначается β. Величина $1 - \beta$ (*мощность критерия*) характеризует вероятность, с которой правильно отвергается неверная гипотеза H_0. Между уровнем значимости α и мощностью критерия $1 - \beta$ существует связь: с уменьшением уровня значимости α, а значит, и с уменьшением вероятности появления ошибки первого рода падает мощность критерия. В этом случае он все меньше улавливает различие между нулевой H_0 и альтернативной H_1 гипотезами. Поэтому нельзя беспредельно уменьшать риск ошибки первого рода, так как суждения становятся все менее определенными. При фиксированном объеме выборки *n* и заданном уровне значимости α предпочтительнее более мощный критерий, на основании которого в большей степени выявляется неверная гипотеза H_0.

Всякий критерий согласия, служащий для проверки статистических гипотез, включает в себя переменную критерия ξ, характеризующую величину суммарных расхождений теоретических и эмпирических показателей (параметров). Множество значений переменной ξ (являющейся по своей сути случайной величиной) разбивается на следующие непересекающиеся части:

- *критическая область*, при попадании ξ в которую основная гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1;
- *область принятия гипотезы* H_0.

Обработка экспериментальных данных с помощью любого критерия согласия осуществляется по следующей схеме:

1. Берется один или два ряда наблюдений (одна или две выборки) и по элементам этих рядов по определенным формулам вычисляют переменную критерия ξ.

2. По заданному уровню значимости α и числу степеней свободы *k* находят по таблицам приложений критическую область значений ξ.

3. Если полученная в пункте 1 переменная ξ попадает в критическую область, то гипотеза H_0 отвергается, принимается конкурирующая гипотеза H_1.

4. Если при проверке гипотезы найденная по выборочным данным ξ не попадает в критическую область, то нет достаточных оснований отвергнуть выдвинутую гипотезу H_0. То есть гипотеза H_0 не отвергается, но это не означает, что H_0 является единственно подходящей гипотезой: просто H_0 не противоречит результатам выборочного наблюдения, однако таким же свойством могут обладать наряду с H_0 и другие гипотезы.

4.1. Оценка статистической значимости коэффициента корреляции

Проверяется следующая основная гипотеза $H_0: r_{\text{ген}} = 0$ (о равенстве нулю генерального коэффициента корреляции, то есть об отсутствии линейной зависимости между признаками *X* и *Y* в генеральной совокупности) при конкурирующей гипотезе $H_1: r_{\text{ген}} \neq 0$. При большом объеме выборки *n*, отобранной из генеральной нормально распределенной совокупности, статистическая значимость коэффициента корреляции r_{XY} проверяется на основе *критерия Стьюдента*. В качестве переменной критерия принимается величина

$$\xi = \frac{|r_{XY}|\sqrt{n-2}}{\sqrt{1-r_{XY}^2}},$$

имеющая распределение Стьюдента с $k = n-2$ степенями свободы.

Применение критерия Стьюдента к проверке гипотезы H_0 о равенстве нулю генерального коэффициента корреляции осуществляется по следующему правилу:

1. По найденному выборочному коэффициенту корреляции r_{XY} вычислить величину $\xi_{\text{расч}}$:

$$\xi_{\text{расч}} = \frac{|r_{XY}|\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}.$$

2. По заданному уровню значимости $\alpha = 1-\gamma$ и числу степеней свободы $k = n-2$ определить по таблице распределения Стьюдента (прил. 2) критическое значение $t_{\text{кр}}$.

3. Если $\xi_{\text{расч}} > t_{\text{кр}}$, то гипотеза H_0 отвергается, принимается конкурирующая гипотеза H_1 и, следовательно, полученный коэффициент корреляции r_{XY} *статистически значим*.

4. В случае $\xi_{\text{расч}} \le t_{\text{кр}}$ нет достаточных оснований отвергнуть выдвинутую гипотезу H_0 и, следовательно, данный коэффициент корреляции r_{XY} *статистически незначим*, а полученная линейная регрессия *не может быть* использована как статистическая модель исследуемой взаимосвязи между признаками *X* и *Y*.

Если коэффициент корреляции по модулю близок к единице ($r_{XY} > 0{,}8$), получен по данным относительно малой выборки ($n < 30$), то проверка статистической значимости производится с помощью распределения *Фишера*:

$$z = \frac{1}{2}\ln\frac{1+r_{XY}}{1-r_{XY}}.$$

Средняя квадратическая ошибка распределения Фишера:

$$\sigma(z) = \frac{1}{\sqrt{n-3}}.$$

Если отношение $|z|/\sigma(z) > 3$ (при уровне значимости $\alpha = 0{,}01$), то найденный коэффициент корреляции *статистически значим*; в противном случае — *статистически незначим*. Величина z определяется по таблице значений распределения Фишера.

Задача 1

По выборке объема $n=60$, извлеченной из двумерной нормальной генеральной совокупности (X,Y), найден коэффициент корреляции $r_{XY}=0,747$. Проверить статистическую значимость с вероятностью $\gamma=0,95$ найденного коэффициента корреляции r_{XY}.

Решение

Поскольку $n=60$, $r_{XY}=0,747$, то имеем:

$$\xi_{\text{расч}}=\frac{|r_{XY}|\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}=\frac{0,747\sqrt{60-2}}{\sqrt{1-(0,747)^2}}=8,557\,.$$

Заданный уровень значимости $\alpha=1-\gamma=1-0,95=0,05$, число степеней свободы $k=n-2=60-2=58$. Тогда по таблице распределения Стьюдента (прил. 2) находим $t_{\text{кр}}=2,0$. Поскольку $\xi_{\text{расч}}>t_{\text{кр}}$, то найденный коэффициент корреляции статистически значим (с вероятностью $\gamma=0,95$), то есть X и Y коррелированны.

Задача 2

По выборке объема $n=10$, извлеченной из двумерной нормальной генеральной совокупности (X,Y), найден коэффициент корреляции $r_{XY}=-0,926$. Проверить статистическую значимость найденного коэффициента корреляции r_{XY} при уровне значимости $\alpha=0,01$.

Решение

Поскольку выборка малая ($n=10$, то есть $n<30$), $|r_{XY}|=0,926>0,8$, то значимость проверяем с помощью *распределения Фишера*. По таблице значений распределения Фишера (прил. 3) для $r_{XY}=-0,926$ имеем $z=-1,658$.

Средняя квадратическая ошибка распределения Фишера:

$$\sigma(z)=\frac{1}{\sqrt{n-3}}=\frac{1}{\sqrt{10-3}}=0,378.$$

Вычислим отношение $|z|/\sigma(z)=1,658/0,378=4,387$.

Заданный уровень значимости $\alpha=0,01$. Поскольку $|z|/\sigma(z)=4,387$, а это больше 3, то найденный коэффициент корреляции статистически значим.

4.2. Статистическая проверка гипотезы о теоретическом распределении

Выбор теоретического распределения в качестве математической модели, выражающей закономерность распределения изучаемого признака, производится на основе графических изображений данного вариационного ряда (полигон, гистограмма) и вычисленных статистических показателей (среднее, среднее квадратическое отклонение σ), характеризующих форму и тип кривой распределения. В результате устанавливают конкретный закон распределения, представимый аналитически через функцию распределения $F(x)$ или функцию плотности $f(x)$.

На следующем этапе анализа статистических данных необходимо произвести проверку соответствия найденного теоретического распределения эмпирическому.

При большом объеме выборки $n \geq 100$ и больших частотах $n_i \geq 5$ вариант признака/гипотеза о соответствии найденного теоретического распределения эмпирическому проверяется на основе *критерия согласия* χ^2 *Пирсона*. В качестве переменной критерия χ^2 принимается мера расхождения наблюдаемых частот n_i и теоретических частот n_i':

$$\chi^2 = \sum_{i=1}^{s} \frac{\left(n_i - n_i'\right)^2}{n_i'},$$

имеющая распределение χ^2 с k степенями свободы.

Применение критерия χ^2 к проверке гипотезы о соответствии найденного теоретического распределения эмпирическому осуществляется по следующему правилу:

1. Разбить числовую ось на s промежутков: $(-\infty; a_1)$, $[a_1; a_2)$, ..., $[a_{s-1}, +\infty)$.

2. Рассчитать теоретические частоты n_i', определяющие возможные численности каждого промежутка данного вариационного ряда при предположении о распределении признака по выбранному теоретическому закону $f(x)$:

$$n_i' = nP_i,\ i = 1, 2, \ldots, s,$$

где n — объем совокупности, $n = \sum_i n_i$; $P_i = P\left(a_{i-1} \leq x < a_i\right)$ — вероятность попадания в интервал $\left(a_{i-1}; a_i\right)$ значений признака. В частности, если в качестве теоретического выбрано нормальное распределение с параметрами $\left(a, \sigma^2\right)$, то $P_i = \Phi(u_i) - \Phi(u_{i-1})$, где $\Phi(u)$ — функция Лапласа; u_i — границы интервала $\left(u_i = \dfrac{a_i - a}{\sigma}\right)$.

Если изучаемый признак X является дискретным и x_1, x_2, ..., x_s — его наблюдаемые значения, то $P_i = P(X = x_i)$, $i = 1, 2, ..., s$. Например, при выборе в качестве теоретического распределение Пуассона с параметром λ имеем

$$P_i = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!},$$

где x_i — неотрицательные числа.

3. Вычислить величину переменной критерия:

$$\chi^2_{\text{расч}} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}.$$

Сравнивая наблюдаемые n_i и n_i' теоретические частоты, получаем, что близость их значений говорит в пользу гипотезы о распределении признака в генеральной совокупности по теоретическому закону $f(x)$, заметные различия отвергают эту гипотезу.

4. Определить число степеней свободы *k* по формуле

$$k = s - l - 1,$$

где *s* — число интервалов; *l* — число параметров теоретического закона $f(x)$, вычисляемых с помощью эмпирического распределения, в частности:

- для *нормального распределения* $k = s - 3$, так как по вариационному ряду рассчитываются два параметра:

$$a = \bar{x}, \ \sigma^2 = \sigma^2(X);$$

- для *распределения Пуассона* $k = s - 2$, так как по вариационному ряду рассчитывается один параметр $\lambda = \bar{x}$.

5. По заданному уровню значимости α и числу степеней свободы *k* определить по таблице распределения χ^2 (прил. 4) критическое значение $\chi^2_{\text{кр}}$.

6. Если $\chi^2_{\text{расч}} > \chi^2_{\text{кр}}$, то есть χ^2 попадает в критическую область, то оцениваемое расхождение наблюдаемых частот n_i и теоретических частот n_i' существенно и его нельзя объяснить случайностью выбранных данных. Тогда гипотеза о распределении признака в генеральной совокупности по теоретическому закону $f(x)$ *отвергается*.

7. В случае $\chi^2_{\text{расч}} \le \chi^2_{\text{кр}}$, то оцениваемое расхождение несущественно и может быть объяснено случайностью выбранных данных. В такой ситуации ги-

потеза о распределении признака в генеральной совокупности по теоретическому закону $f(x)$ принимается (не отвергается).

4.3. Построение кривой распределения по эмпирическим данным. Проверка гипотезы о нормальном распределении выборки

В данном разделе приведена методика проведения лабораторной работы по решению основной задачи математической статистики — определение вида распределения изучаемого признака на основе экспериментальных данных, проверка согласованности гипотезы о распределении генеральной совокупности с полученным выборочным распределением.

Цель работы:

- овладение способами построения эмпирической и теоретической кривой распределения;
- выработка умения и навыков применения критерия согласия Пирсона для проверки выдвинутой статистической гипотезы.

Задача 3

Произведено статистическое наблюдение признака *X*, результаты которого приведены в таблице.

Интервалы значений признака	3,0–3,6	3,6–4,2	4,2–4,8	4,8–5,4	5,5–6,0	6,0–6,6	6,6–7,2
Частота	5	14	38	55	27	15	6

Построить эмпирическую и теоретическую кривую распределения.

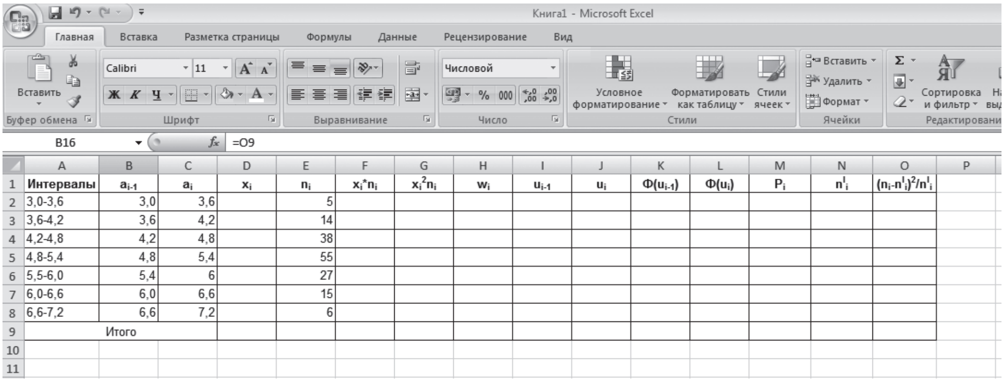
При уровне значимости $\alpha = 0{,}01$ проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности с полученным выборочным распределением.

Порядок выполнения работы

Запуск Microsoft Excel

Запустить **Microsoft Excel**, в новом документе удалить листы 2 и 3, сохранить файл в своей папке под именем «Задача 4». Далее необходимо периодически сохранять документ.

Переименовать *Лист1* в *Расчетная таблица*. Расчетная таблица должна иметь следующий вид (рис. 4.1).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**Интервалы**	**a_{i-1}**	**a_i**	**x_i**	**n_i**	**x_i*n_i**	**$x_i^2n_i$**	**w_i**	**u_{i-1}**	**u_i**	**$\Phi(u_{i-1})$**	**$\Phi(u_i)$**	**P_i**	**n'_i**	**$(n_i-n'_i)^2/n'_i$**
2	3,0-3,6	3,0	3,6		5										
3	3,6-4,2	3,6	4,2		14										
4	4,2-4,8	4,2	4,8		38										
5	4,8-5,4	4,8	5,4		55										
6	5,5-6,0	5,4	6		27										
7	6,0-6,6	6,0	6,6		15										
8	6,6-7,2	6,6	7,2		6										
9	Итого														

Рис. 4.1. Расчетная таблица

Заполнение расчетной таблицы

Набрать заголовки в первой строке. Чтобы сделать текст нижним или верхним индексом, можно действовать так: сначала набрать в ячейке B1 просто a_{i-1}, затем в строке формул выделить $i-1$, выбрать на вкладке *Главная Шрифт*, там, в разделе *Видоизменение*, установить галочку напротив надписи «подстрочный». Аналогично делаются индексы в остальных ячейках. Сделать заголовки полужирным, выравнивание по центру.

Ввести данные, известные из условия задачи: частоты n_i, левые границы интервалов a_{i-1}, правые границы a_i.

В столбце D необходимо вычислить середины интервалов x_i по формуле

$$x_i = \frac{a_{i-1} + a_i}{2}.$$

В ячейку D2 ввести следующую формулу:

$$=(B2+C2)/2.$$

В остальные ячейки столбца D копируйте формулы с помощью маркера заполнения.

Далее заполните столбцы F и G по формулам

$$x_i n_i = x_i \cdot n_i,$$

$$x_i^2 n_i = x_i \cdot x_i \cdot n_i$$

соответственно. Для этого в ячейку F2 ввести следующую расчетную формулу:

$$=D2*E2,$$

в ячейку G2 — формулу

$$=D2*D2*E2.$$

Остальные ячейки столбцов F и G заполнить с помощью маркера заполнения.

В строке *Итого* в столбцах E, F и G вычислить итоговые суммы. Для этого установить курсор мыши на ячейке E9 и нажать кнопку «*Автосумма*» на панели инструментов вкладки *Формула*. Ячейки F9 и G9 заполнить с помощью маркера заполнения.

Вычисление числовых характеристик статистической выборки

В ячейке A12 сделать подпись $x_{\text{ср}} =$. *Выборочное среднее* — среднее арифметическое значений выборки вычисляется по формуле

$$x_{\text{ср}} = \frac{\sum x_i n_i}{\sum n_i}.$$

Суммарные значения произведений $x_i n_i$ мы вычислили в ячейке F9, суммарные значения n_i — в ячейке E9. Таким образом, в ячейку B12 ввести следующую формулу:

$$=F9/E9.$$

Аналогично в строке 13, в ячейке A13 ввести $x^2{}_{\text{ср}} =$ и вычислить среднее арифметическое квадратов значений выборки $x^2{}_{\text{ср}}$ по формуле

$$x^2{}_{\text{ср}} = \frac{\sum x_i^2 n_i}{\sum n_i}.$$

Следующую строку отвести под *выборочную дисперсию* $D_{\text{В}}$ — среднее значение квадрата отклонения $x_i - x_{\text{ср}}$, которую необходимо вычислить по формуле

$$D_{\text{В}} = x^2{}_{\text{ср}} - (x_{\text{ср}})^2.$$

В формулах Excel для операции возведения в степень используется символ ^ (он получается нажатием одновременно клавиш «Shift» и «6»). Итак, в ячейке A14 сделать подпись $D(X) =$, в ячейку B14 ввести формулу

$$=B13-B12\,\hat{}\,2.$$

В следующей строке выборочное среднее квадратическое отклонение $\sigma_В$ необходимо подписать как SI (*X*) (ячейка A15) и вычислить значение по формуле

$$\sigma_В = \sqrt{D_В}\ .$$

Для вычисления корня перейти в ячейку B15, на панели инструментов вкладки *Формула* нажмите кнопку «*Вставить функцию* f_x». В открывшемся окне *Мастера функций* выбрать в списке категорий *Математические*, в списке функций, используя полосу прокрутки, выбрать *КОРЕНЬ*, щелкнуть *OK,* далее в окне *Аргументы функции* в поле *Число* ввести адрес ячейки B14, нажать *OK*.

Заполнить столбец *относительных частот* w_i согласно формуле

$$w_i = \frac{n_i}{\sum n_i}\ .$$

В ячейку H2 ввести формулу

$$= E2 / \$E\$9\ .$$

Для проверки подсчитать итоговую сумму в столбце H, которая должна быть равна 1.

На данном этапе выполнения работы должна получиться расчетная таблица, как показано на рис. 4.2.

Построение гистограммы относительных частот вариационного ряда

Построим гистограмму относительных частот вариационного ряда.

На вкладке *Вставка* в строке меню *Диаграммы* выбрать вид диаграммы *Гистограмма,* затем подвид диаграммы *Гистограмма с группировкой* (слева в первом ряду).

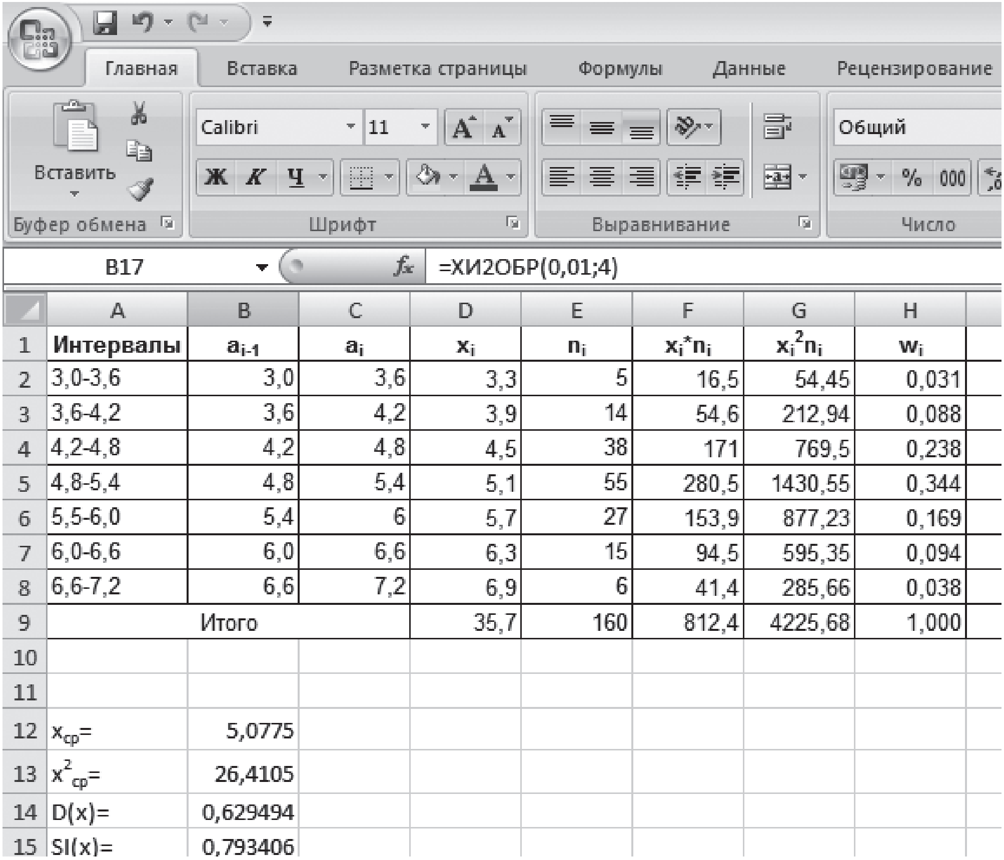
В результате на рабочем листе появится окно с пустой диаграммой. Для отображения на ней данных выполнить следующие действия:

1. В контекстном меню диаграммы выполнить пункт *Выбрать данные*.

2. В открывшемся окне в поле *Диапазон данных для диаграммы* указать ячейки, в которых находятся относительные частоты w_i. В разделе *Подписи горизонтальной оси* нажать *Изменить* и указать в поле *Диапазон подписей оси* ячейки, в которых находятся средины интервалов x_i. Нажать *OK*.

Редактирование гистограммы

Щелкнуть по диаграмме, на вкладке *Макет* в группе *Фон* нажать на кнопку *Область построения*, выбрать в выпадающем меню вариант *Нет* (*удалить заливку области построения*).



	A	B	C	D	E	F	G	H
1	**Интервалы**	a_{i-1}	a_i	x_i	n_i	x_i*n_i	$x_i^2 n_i$	w_i
2	3,0-3,6	3,0	3,6	3,3	5	16,5	54,45	0,031
3	3,6-4,2	3,6	4,2	3,9	14	54,6	212,94	0,088
4	4,2-4,8	4,2	4,8	4,5	38	171	769,5	0,238
5	4,8-5,4	4,8	5,4	5,1	55	280,5	1430,55	0,344
6	5,5-6,0	5,4	6	5,7	27	153,9	877,23	0,169
7	6,0-6,6	6,0	6,6	6,3	15	94,5	595,35	0,094
8	6,6-7,2	6,6	7,2	6,9	6	41,4	285,66	0,038
9		Итого		35,7	160	812,4	4225,68	1,000
10								
11								
12	$x_{ср}$=	5,0775						
13	$x^2_{ср}$=	26,4105						
14	D(x)=	0,629494						
15	SI(x)=	0,793406						

Рис. 4.2. Промежуточные результаты расчетной таблицы

Нажать правую клавишу мыши на одном из столбцов построенной гистограммы, выбрать команду *Формат ряда данных*, пункт *Параметры ряда*, там установить *Боковой зазор* 0 %, в пункте *Заливка* установить флажок *Разноцветные точки*.

Щелкнуть кнопку *Названия осей*, далее выбрать *Название основной горизонтальной оси*, затем вариант *Название под осью*. Ввести вместо *Названия оси Интервалы*.

Поместить диаграмму на *отдельном листе* с именем *Гистограмма*. Для этого щелкнуть диаграмму, на вкладке *Конструктор* в группе *Расположение* нажать кнопку *Переместить диаграмму*. В появившемся окне в разделе *Разместить диаграмму* установить переключатель в положение *На отдельном листе* и ввести имя нового листа диаграммы в поле этого переключателя. Нажать *ОК*.

На рис. 4.3 приведена гистограмма, которую необходимо получить.

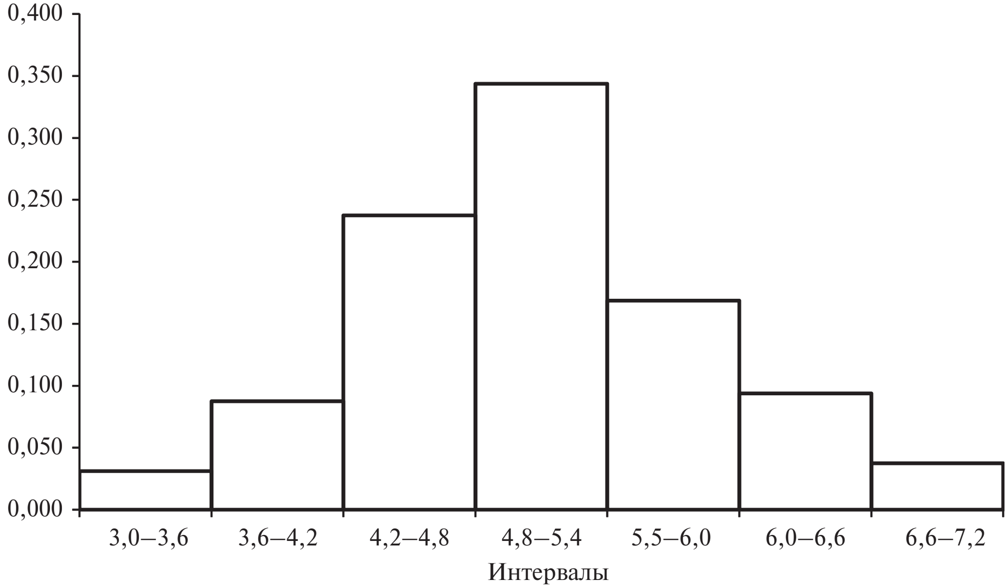


Рис. 4.3. Гистограмма

Проверка согласованности эмпирического распределения с теоретическим нормальным распределением

По виду эмпирической кривой распределения (рис. 4.3) можно сделать предположение, что основная гипотеза H_0 о нормальном распределении генеральной совокупности будет принята.

Нормальное распределение полностью определяется двумя неизвестными параметрами a и σ^2, поэтому, прежде всего, необходимо вычислить точечные оценки, a^*, σ^{2*} этих параметров:

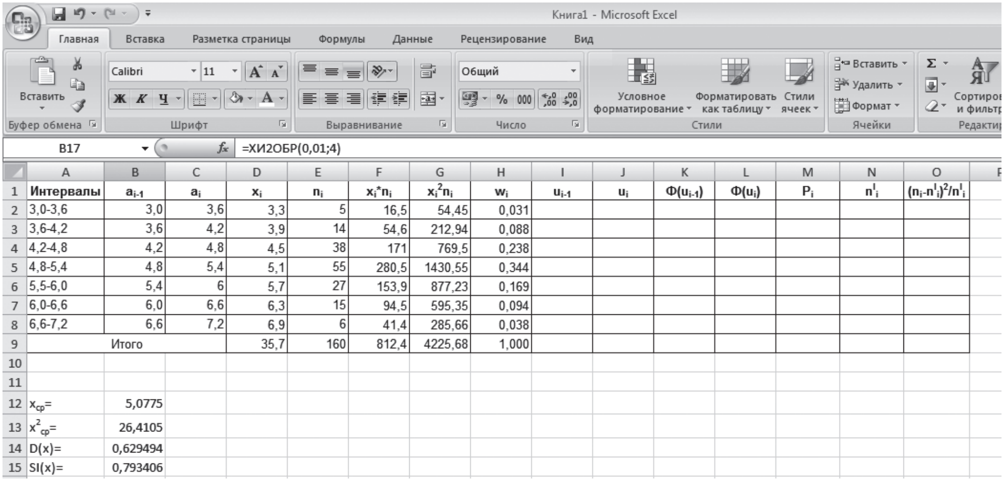
$$a^* = \bar{x} = x_{\text{ср}} = 5{,}0775,$$

$$\sigma^{2*} = \sigma_{\text{В}}{}^2 = D(X) = 0{,}6295.$$

Таким образом, точечные оценки параметров a^*, σ^{2*} нормального распределения уже посчитаны в ячейках B12 и B14 расчетной таблицы соответственно. То есть в качестве теоретического закона выбирается следующая функция плотности:

$$f(x) = \frac{1}{\sqrt{2\pi 0{,}6295}} e^{-\frac{(x-5{,}0775)^2}{2\cdot 0{,}6295}}.$$

Для проверки гипотезы о таком предполагаемом теоретическом законе добавить к расчетной таблице дополнительные столбцы по образцу, представленному на рис. 4.4.



B17 =ХИ2ОБР(0,01;4)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Интервалы	a_{i-1}	a_i	x_i	n_i	x_i*n_i	$x_i^2n_i$	w_i	u_{i-1}	u_i	$\Phi(u_{i-1})$	$\Phi(u_i)$	P_i	n^I_i	$(n_i-n^I_i)^2/n^I_i$
2	3,0-3,6	3,0	3,6	3,3	5	16,5	54,45	0,031							
3	3,6-4,2	3,6	4,2	3,9	14	54,6	212,94	0,088							
4	4,2-4,8	4,2	4,8	4,5	38	171	769,5	0,238							
5	4,8-5,4	4,8	5,4	5,1	55	280,5	1430,55	0,344							
6	5,5-6,0	5,4	6	5,7	27	153,9	877,23	0,169							
7	6,0-6,6	6,0	6,6	6,3	15	94,5	595,35	0,094							
8	6,6-7,2	6,6	7,2	6,9	6	41,4	285,66	0,038							
9		Итого		35,7	160	812,4	4225,68	1,000							
10															
11															
12	$x_{ср}$=	5,0775													
13	$x^2_{ср}$=	26,4105													
14	D(x)=	0,629494													
15	SI(x)=	0,793406													

Рис. 4.4. Дополнительные столбцы расчетной таблицы

В столбце I необходимо вычислить левые концы интервалов u_{i-1} по формуле

$$u_{i-1} = \frac{a_{i-1} - x_{\text{ср}}}{\sigma_{\text{В}}}.$$

В ячейку I2 ввести $-\infty$. В ячейку I3 ввести следующую формулу:

$$=(B3-\$B\$12)/\$B\$15.$$

В остальные ячейки столбца I скопировать формулы с помощью маркера заполнения.

В столбце J необходимо вычислить правые концы интервалов u_i по формуле

$$u_i = \frac{a_i - x_{\text{ср}}}{\sigma_{\text{В}}}.$$

В ячейку J2 введите следующую формулу:

$$=(C2-\$B\$12)/\$B\$15.$$

В диапазон ячеек J3-J7 столбца J скопировать формулы с помощью маркера заполнения. В ячейку J8 ввести $+\infty$.

Далее заполнить столбцы K и L по таблице *значений функции Лапласа* (прил. 1), причем $\Phi(-\infty)=-0{,}5$, $\Phi(+\infty)=0{,}5$.

Заполнить столбец M вероятностей P_i по формуле

$$P_i=\Phi(u_i)-\Phi(u_{i-1}).$$

В ячейку M2 ввести следующую формулу:

$$=(L2-K2).$$

В остальные ячейки столбца M скопировать формулы с помощью маркера заполнения. В ячейке M9 вычислить сумму вероятностей: она должна равняться *единице*. Проверьте!

Далее заполнить столбец N теоретических частот n_i' по формуле

$$n_i'=nP_i.$$

В ячейку N2 ввести следующую формулу:

$$=\$E\$9*M2.$$

В остальные ячейки столбца N скопировать формулы с помощью маркера заполнения. В ячейке N9 вычислить сумму теоретических частот n_i'.

Заполнить столбец O — величина отношения квадрата отклонения теоретической частоты от эмпирической к теоретической частоте: $\dfrac{(n_i-n_i')^2}{n_i'}$.

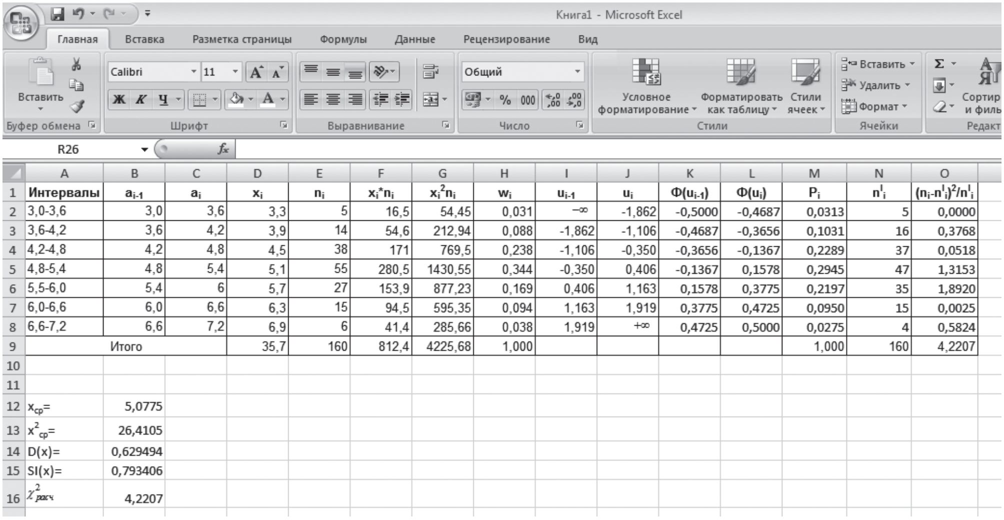
В ячейку O2 ввести следующую формулу:

$$=\frac{(E2-N2)\hat{}2}{N2}.$$

В остальные ячейки столбца O скопировать формулы с помощью маркера заполнения. В ячейке O9 вычислить сумму теоретических частот отношения $\dfrac{(n_i-n_i')^2}{n_i'}$, то есть переменную критерия $\chi^2_{\text{расч}}$.

В ячейке A16 ввести $\chi^2_{\text{расч}}$, в ячейку B16 скопировать значения ячейки O9.

Окончательно должны получиться результаты, как на рис. 4.5.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Интервалы	a_{i-1}	a_i	x_i	n_i	x_i*n_i	$x_i^2 n_i$	w_i	u_{i-1}	u_i	$\Phi(u_{i-1})$	$\Phi(u_i)$	P_i	n'_i	$(n_i-n'_i)^2/n'_i$
2	3,0-3,6	3,0	3,6	3,3	5	16,5	54,45	0,031	$-\infty$	-1,862	-0,5000	-0,4687	0,0313	5	0,0000
3	3,6-4,2	3,6	4,2	3,9	14	54,6	212,94	0,088	-1,862	-1,106	-0,4687	-0,3656	0,1031	16	0,3768
4	4,2-4,8	4,2	4,8	4,5	38	171	769,5	0,238	-1,106	-0,350	-0,3656	-0,1367	0,2289	37	0,0518
5	4,8-5,4	4,8	5,4	5,1	55	280,5	1430,55	0,344	-0,350	0,406	-0,1367	0,1578	0,2945	47	1,3153
6	5,5-6,0	5,4	6	5,7	27	153,9	877,23	0,169	0,406	1,163	0,1578	0,3775	0,2197	35	1,8920
7	6,0-6,6	6,0	6,6	6,3	15	94,5	595,35	0,094	1,163	1,919	0,3775	0,4725	0,0950	15	0,0025
8	6,6-7,2	6,6	7,2	6,9	6	41,4	285,66	0,038	1,919	$+\infty$	0,4725	0,5000	0,0275	4	0,5824
9	Итого			35,7	160	812,4	4225,68	1,000					1,000	160	4,2207
10															
11															
12	$x_{ср}$=	5,0775													
13	$x^2_{ср}$=	26,4105													
14	D(x)=	0,629494													
15	SI(x)=	0,793406													
16	$\chi^2_{расч}$	4,2207													

Рис. 4.5. Таблица результатов

Определим число степеней свободы $k = s - 3 = 7 - 3 = 4$. При уровне значимости $\alpha = 0{,}01$ и числе степеней свободы $k = 4$ найдем табличное значение $\chi^2_{кр} = 13{,}2767$ (прил. 4). В то же время $\chi^2_{расч} = 4{,}2207$ (см. рис. 45), то есть $\chi^2_{расч} < \chi^2_{кр}$, и, следовательно, гипотеза о нормальном распределении генеральной совокупности с выбранным теоретическим $f(x)$ законом *принимается*.

Построение эмпирической и теоретической кривых распределения

Необходимо построить *эмпирическую* и *теоретические* кривые распределения, то есть кривую *частот* n_i вариационного ряда и *теоретических частот* n'_i.

На вкладке *Вставка* в строке меню *Диаграммы* выбрать вид диаграммы *Точечная,* затем подвид диаграммы *Точечная с прямыми отрезками*.

В результате на рабочем листе появится окно с пустой диаграммой. Для отображения на ней данных выполнить следующие действия:

1. В контекстном меню диаграммы выполнить пункт *Выбрать данные*.

2. В открывшемся окне в поле *Диапазон данных для диаграммы* указать ячейки, в которых находятся эмпирические частоты n_i, ячейки, в которых находятся теоретические частоты n'_i, и ячейки, в которых находятся середины интервалов x_i. Нажать *OK*.

Редактирование кривых

Расположить диаграмму на *отдельном листе* с именем *Кривые распределения*. Для этого щелкнуть диаграмму, вызвать контекстное меню и выполнить пункт *Переместить диаграмму*. В появившемся окне в разделе *Разместить* *диаграмму* установить переключатель в положение *На отдельном листе* и ввести имя нового листа диаграммы в поле этого переключателя. Нажать *OK*.

Щелкнуть по диаграмме, в командной строке *Работа с диаграммами* на вкладке *Макет* в группе *Фон* нажать на кнопку *Область построения*, выбрать в выпадающем меню вариант *Нет* (*удалить заливку области построения*).

В группе *Оси* кнопки *Оси* и *Сетка* позволяют настроить отображение осей и линий сетки для диаграммы. Опции *Основная горизонтальная ось* и *Основная вертикальная ось* выбираются по умолчанию. Создать и горизонтальные, и вертикальные линии сетки для основных делений — опция *Основные линии сетки*.

Вызвать контекстное меню (щелкните правой кнопкой мыши) на линии теоретических частот, выбрать команду *Формат ряда данных*, пункт *Тип линии*, в подпункте *Тип штриха* установить *Штрих*.

Для настройки параметров горизонтальной и вертикальной осей (форматирование этих элементов диаграммы) следует перейти к окну *Формат оси* с помощью контекстного меню. Для *горизонтальной* оси выбрать на вкладке *Параметры оси* минимальное значение *3,3*; максимальное значение *6,9*; цену основных делений *0,6*; для *вертикальной* оси — минимальное значение *0*, максимальное значение *60*, цену основных делений *10*.

Щелкнуть по диаграмме, в командной строке *Работа с диаграммами* на вкладке *Макет* в группе команд *Подписи* нажать кнопку *Названия осей*, выбрать *Название основной горизонтальной оси*, затем вариант *Название под осью*. Ввести вместо *Названия оси* обозначение x, затем перетащить его (методом «буксировки») в положение справа от оси. Выполнить аналогичные действия для названия основной вертикальной оси, введя обозначения для нее *Частота* и перетащив его в положение над осью. Щелкнув кнопку *Легенда*, выбрать вариант *Нет*.

На рис. 4.6 приведены кривые, которые необходимо получить.

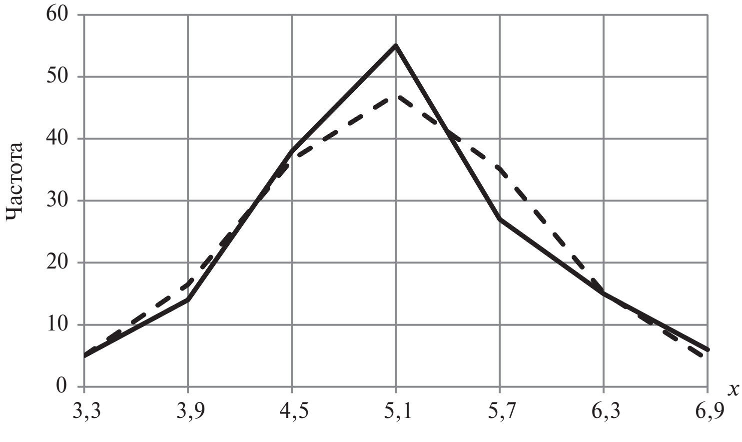


Рис. 4.6. Кривые распределения:
сплошная — эмпирическая; штриховая — теоретическая

Индивидуальное задание

Создайть лист *Расчетная таблица №*. Решить свой вариант задания.

Результаты работы

В результате выполненной лабораторной работы студент должен продемонстрировать преподавателю готовый файл *Задача 4.xlsx*, содержащий 7 листов:

- лист 1 (*Титульный лист*) — титульный лист к работе, на котором указаны название работы, номер варианта, Ф. И.О. студента, номер группы, дата выполнения работы;
- лист 2 (*Расчетная таблица*) — таблица исходных данных и решения задачи;
- лист 3 (*Гистограмма*) — гистограмма задачи;
- лист 4 (*Кривые распределения*) — графики теоретической и эмпирической кривых распределения задачи;
- лист 5 (*Расчетная таблица №...*) — таблица исходных данных и решения задания своего варианта;
- лист 6 (*Гистограмма №...*) — гистограмма задания своего варианта;
- лист 7 (*Кривые распределения №...*) — графики теоретической и эмпирической кривых распределения задания своего варианта.

Варианты индивидуального задания

Вариант 1

При уровне значимости $\alpha = 0,05$ проверить гипотезу о нормальном распределении генеральной совокупности при следующем эмпирическом распределении выборки объемом $n = 100$.

Интервалы значений признака	3–8	8–13	13–18	18–23	23–28	28–33	33–38
Частота	6	8	15	40	16	8	7

Построить эмпирическую и теоретическую кривую распределения.

Вариант 2

При уровне значимости $\alpha = 0,01$ проверить, согласуется ли гипотеза о нормальном распределении исследуемого признака *X* в генеральной совокупности с полученным выборочным распределением.

Интервалы значений признака	1–6	6–11	11–16	16–21
Частота	12	28	47	13

Построить эмпирическую и теоретическую кривую распределения.

Вариант 3

При уровне значимости $\alpha = 0{,}05$ проверить гипотезу о нормальном распределении генеральной совокупности при следующем эмпирическом распределении выборки объемом $n = 300$.

Интервалы значений признака	-20– -10	-10–0	0–10	10–20	20–30	30–40	40–50
Частота	20	47	80	89	40	16	8

Построить эмпирическую и теоретическую кривую распределения.

Вариант 4

При уровне значимости $\alpha = 0{,}01$ проверить, согласуется ли гипотеза о нормальном распределении исследуемого признака X в генеральной совокупности с полученным выборочным распределением.

Интервалы значений признака	20–30	30–40	40–50	50–60	60–70	70–80
Частота	8	19	28	32	42	21

Построить эмпирическую и теоретическую кривую распределения.

Вариант 5

При уровне значимости $\alpha = 0{,}05$ проверить гипотезу о нормальном распределении генеральной совокупности при следующем эмпирическом распределении выборки объемом $n = 200$.

Интервалы значений признака	0–5	5–10	10–15	15–20	20–25	25–30
Частота	20	30	40	50	40	20

Построить эмпирическую и теоретическую кривую распределения.

Вариант 6

При уровне значимости $\alpha = 0{,}01$ проверить, согласуется ли гипотеза о нормальном распределении исследуемого признака X в генеральной совокупности с полученным выборочным распределением.

Интервалы значений признака	До 5	5–10	10–15	15–20	20–25	25–30	30 и более
Частота	8	95	204	270	210	130	83

Построить эмпирическую и теоретическую кривую распределения.

Вариант 7

При уровне значимости $\alpha = 0,05$ проверить гипотезу о нормальном распределении генеральной совокупности при следующем эмпирическом распределении выборки объемом $n = 144$.

Интервалы значений признака	До 2	2–4	4–6	6–8	8–10	10–12	Более 12
Частота	26	30	28	24	18	10	8

Построить эмпирическую и теоретическую кривую распределения.

Вариант 8

При уровне значимости $\alpha = 0,01$ проверить, согласуется ли гипотеза о нормальном распределении исследуемого признака X в генеральной совокупности с полученным выборочным распределением.

Интервалы значений признака	80–100	100–120	120–140	140–160	160–180
Частота	10	20	40	20	10

Построить эмпирическую и теоретическую кривую распределения.

Вариант 9

При уровне значимости $\alpha = 0,05$ проверить гипотезу о нормальном распределении генеральной совокупности при следующем эмпирическом распределении выборки объемом $n = 100$.

Интервалы значений признака	158–162	162–166	166–170	170–174	174–178	178–182	182–186
Частота	10	14	26	28	12	8	2

Построить эмпирическую и теоретическую кривую распределения.

Вариант 10

При уровне значимости $\alpha = 0,01$ проверить, согласуется ли гипотеза о нормальном распределении исследуемого признака X в генеральной совокупности с полученным выборочным распределением.

Интервалы значений признака	1–5	5–9	9–13	13–17	17–21
Частота	10	20	50	12	8

Построить эмпирическую и теоретическую кривую распределения.

Рекомендуемый библиографический список

Айвазян С. А. Прикладная статистика: Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. М. : Финансы и статистика, 2012. 323 с.

Вентцель Е. С. Теория случайных процессов и ее инженерные приложения / Е. С. Вентцель, Л. А. Овчаров. М. : Высшая школа, 2011. 287 с.

Глушаков С. В. Microsoft Excel 2007. Лучший самоучитель / С. В. Глушаков, А. С. Сурядный. М. : АСТ, 2012. 364 с.

Гмурман В. Е. Теория вероятностей и математическая статистика: учеб. пособие для вузов / В. Е. Гмурман. 12-е изд., перераб. М. : Юрайт, 2011. 479 с.

Горелова Г. В. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel / Г. В. Горелова, И. А. Кацко. Ростов н/Д. : Феникс, 2010. 257 с.

Гребенникова И. В. Математика в юриспруденции / И. В. Гребенникова, А. Г. Кремлев, Е. А. Залазинская. Екатеринбург : Издательский дом УрГЮА, 2004. 164 с.

Севастьянов Б. А. Курс теории вероятностей и математическая статистика / Б. А. Севастьянов. М. : Институт компьютерных исследований, 2010. 372 с.

Чудесенко В. Ф. Сборник заданий по спецкурсам высшей математики (типовые расчеты) / В. Ф. Чудесенко. М. : Лань, 2012. 265 с.

Чураков Е. П. Математические методы обработки экспериментальных данных в экономике : учеб. пособие / Е. П. Чураков. М. : Финансы и статистика, 2011. 240 с.

Приложение 1

Таблица значений функции Лапласа

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,35	0,1368	0,70	0,2580	1,05	0,3531	1,40	0,4192
0,01	0,0040	0,36	0,1406	0,71	0,2611	1,06	0,3554	1,41	0,4207
0,02	0,0080	0,37	0,1443	0,72	0,2642	1,07	0,3577	1,42	0,4222
0,03	0,0120	0,38	0,1480	0,73	0,2673	1,08	0,3599	1,43	0,4236
0,04	0,0160	0,39	0,1517	0,74	0,2704	1,09	0,3621	1,44	0,4251
0,05	0,0199	0,40	0,1554	0,75	0,2734	1,10	0,3643	1,45	0,4265
0,06	0,0239	0,41	0,1591	0,76	0,2764	1,11	0,3665	1,46	0,4279
0,07	0,0279	0,42	0,1628	0,77	0,2794	1,12	0,3686	1,47	0,4292
0,08	0,0319	0,43	0,1664	0,78	0,2823	1,13	0,3708	1,48	0,4306
0,09	0,0359	0,44	0,1700	0,79	0,2852	1,14	0,3729	1,49	0,4319
0,10	0,0398	0,45	0,1736	0,80	0,2881	1,15	0,3749	1,50	0,4332
0,11	0,0438	0,46	0,1772	0,81	0,2910	1,16	0,3770	1,51	0,4345
0,12	0,0478	0,47	0,1808	0,82	0,2939	1,17	0,3790	1,52	0,4357
0,13	0,0517	0,48	0,1844	0,83	0,2967	1,18	0,3810	1,53	0,4370
0,14	0,0557	0,49	0,1879	0,84	0,2995	1,19	0,3830	1,54	0,4382
0,15	0,0596	0,50	0,1915	0,85	0,3023	1,20	0,3849	1,55	0,4394
0,16	0,0636	0,51	0,1950	0,86	0,3051	1,21	0,3869	1,56	0,4406
0,17	0,0675	0,52	0,1985	0,87	0,3078	1,22	0,3888	1,57	0,4418
0,18	0,0714	0,53	0,2019	0,88	0,3106	1,23	0,3907	1,58	0,4429
0,19	0,0753	0,54	0,2054	0,89	0,3133	1,24	0,3925	1,59	0,4441
0,20	0,0793	0,55	0,2088	0,90	0,3159	1,25	0,3944	1,60	0,4452
0,21	0,0832	0,56	0,2123	0,91	0,3186	1,26	0,3962	1,61	0,4463
0,22	0,0871	0,57	0,2157	0,92	0,3212	1,27	0,3980	1,62	0,4474
0,23	0,0910	0,58	0,2190	0,93	0,3238	1,28	0,3997	1,63	0,4484
0,24	0,0948	0,59	0,2224	0,94	0,3264	1,29	0,4015	1,64	0,4495
0,25	0,0987	0,60	0,2257	0,95	0,3289	1,30	0,4032	1,65	0,4505
0,26	0,1026	0,61	0,2291	0,96	0,3315	1,31	0,4049	1,66	0,4515
0,27	0,1064	0,62	0,2324	0,97	0,3340	1,32	0,4066	1,67	0,4525
0,28	0,1103	0,63	0,2357	0,98	0,3365	1,33	0,4082	1,68	0,4535
0,29	0,1141	0,64	0,2389	0,99	0,3389	1,34	0,4099	1,69	0,4545
0,30	0,1179	0,65	0,2422	1,00	0,3413	1,35	0,4115	1,70	0,4554
0,31	0,1217	0,66	0,2454	1,01	0,3438	1,36	0,4131	1,71	0,4564
0,32	0,1255	0,67	0,2486	1,02	0,3461	1,37	0,4147	1,72	0,4573
0,33	0,1293	0,68	0,2517	1,03	0,3485	1,38	0,4162	1,73	0,4582
0,34	0,1331	0,69	0,2549	1,04	0,3508	1,39	0,4177	1,74	0,4591

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,75	0,4599	2,15	0,4842	2,55	0,4946	2,95	0,49841	3,35	0,49960
1,76	0,4608	2,16	0,4846	2,56	0,4948	2,96	0,49846	3,36	0,49961
1,77	0,4616	2,17	0,4850	2,57	0,4949	2,97	0,49851	3,37	0,49962
1,78	0,4625	2,18	0,4854	2,58	0,4951	2,98	0,49856	3,38	0,49964
1,79	0,4633	2,19	0,4857	2,59	0,4952	2,99	0,49861	3,39	0,49965
1,80	0,4641	2,20	0,4861	2,60	0,4953	3,00	0,49865	3,40	0,49966
1,81	0,4649	2,21	0,4864	2,61	0,4955	3,01	0,49869	3,41	0,49968
1,82	0,4656	2,22	0,4868	2,62	0,4956	3,02	0,49874	3,42	0,49969
1,83	0,4664	2,23	0,4871	2,63	0,4957	3,03	0,49878	3,43	0,4997
1,84	0,4671	2,24	0,4875	2,64	0,4959	3,04	0,49882	3,44	0,4997
1,85	0,4678	2,25	0,4878	2,65	0,4960	3,05	0,49886	3,45	0,4997
1,86	0,4686	2,26	0,4881	2,66	0,4961	3,06	0,49889	3,46	0,4997
1,87	0,4693	2,27	0,4884	2,67	0,4962	3,07	0,49893	3,47	0,4997
1,88	0,4699	2,28	0,4887	2,68	0,4963	3,08	0,49896	3,48	0,4997
1,89	0,4706	2,29	0,4890	2,69	0,4964	3,09	0,49900	3,49	0,4998
1,90	0,4713	2,30	0,4893	2,70	0,4965	3,10	0,49903	3,50	0,499767
1,91	0,4719	2,31	0,4896	2,71	0,4966	3,11	0,49906	3,51	0,499776
1,92	0,4726	2,32	0,4898	2,72	0,4967	3,12	0,49910	3,52	0,499784
1,93	0,4732	2,33	0,4901	2,73	0,4968	3,13	0,49913	3,53	0,499792
1,94	0,4738	2,34	0,4904	2,74	0,4969	3,14	0,49916	3,54	0,499800
1,95	0,4744	2,35	0,4906	2,75	0,4970	3,15	0,49918	3,55	0,499807
1,96	0,4750	2,36	0,4909	2,76	0,4971	3,16	0,49921	3,56	0,499815
1,97	0,4756	2,37	0,4911	2,77	0,4972	3,17	0,49924	3,57	0,499822
1,98	0,4761	2,38	0,4913	2,78	0,4973	3,18	0,49926	3,58	0,499828
1,99	0,4767	2,39	0,4916	2,79	0,4974	3,19	0,49929	3,59	0,499835
2,00	0,4772	2,40	0,4918	2,80	0,49744	3,20	0,49931	3,60	0,499841
2,01	0,4778	2,41	0,4920	2,81	0,49752	3,21	0,49934	3,61	0,499847
2,02	0,4783	2,42	0,4922	2,82	0,49760	3,22	0,49936	3,62	0,499853
2,03	0,4788	2,43	0,4925	2,83	0,49767	3,23	0,49938	3,63	0,499858
2,04	0,4793	2,44	0,4927	2,84	0,49774	3,24	0,49940	3,64	0,499864
2,05	0,4798	2,45	0,4929	2,85	0,49781	3,25	0,49942	3,65	0,499869
2,06	0,4803	2,46	0,4931	2,86	0,49788	3,26	0,49944	3,66	0,499874
2,07	0,4808	2,47	0,4932	2,87	0,49795	3,27	0,49946	3,67	0,499879
2,08	0,4812	2,48	0,4934	2,88	0,49801	3,28	0,49948	3,68	0,499883
2,09	0,4817	2,49	0,4936	2,89	0,49807	3,29	0,49950	3,69	0,499888
2,10	0,4821	2,50	0,4938	2,90	0,49813	3,30	0,49952	3,70	0,499892
2,11	0,4826	2,51	0,4940	2,91	0,49819	3,31	0,49953	3,71	0,499896
2,12	0,4830	2,52	0,4941	2,92	0,49825	3,32	0,49955	3,72	0,499900
2,13	0,4834	2,53	0,4943	2,93	0,49831	3,33	0,49957	3,73	0,499904
2,14	0,4838	2,54	0,4945	2,94	0,49836	3,34	0,49958	3,74	0,499908

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
3,75	0,499912	4,16	0,499984	4,57	0,4999976	4,98	0,49999968
3,76	0,499915	4,17	0,499985	4,58	0,4999977	4,99	0,49999970
3,77	0,499918	4,18	0,499985	4,59	0,4999978	5,00	0,49999971
3,78	0,499922	4,19	0,499986	4,60	0,4999979	5,01	0,49999973
3,79	0,499925	4,20	0,499987	4,61	0,4999980	5,02	0,49999974
3,80	0,499928	4,21	0,499987	4,62	0,4999981	5,03	0,49999975
3,81	0,499931	4,22	0,499988	4,63	0,4999982	5,04	0,49999977
3,82	0,499933	4,23	0,4999883	4,64	0,49999826	5,05	0,49999978
3,83	0,499936	4,24	0,4999888	4,65	0,49999834	5,06	0,49999979
3,84	0,499938	4,25	0,4999893	4,66	0,49999842	5,07	0,49999980
3,85	0,499941	4,26	0,4999898	4,67	0,49999849	5,08	0,49999981
3,86	0,499943	4,27	0,4999902	4,68	0,49999857	5,09	0,49999982
3,87	0,499946	4,28	0,4999907	4,69	0,49999863	5,10	0,49999983
3,88	0,499948	4,29	0,4999911	4,70	0,49999870	5,11	0,49999984
3,89	0,499950	4,30	0,4999915	4,71	0,49999876	5,12	0,49999985
3,90	0,499952	4,31	0,4999918	4,72	0,49999882	5,13	0,49999986
3,91	0,499954	4,32	0,4999922	4,73	0,49999888	5,14	0,49999986
3,92	0,499956	4,33	0,4999925	4,74	0,49999893	5,15	0,49999987
3,93	0,499958	4,34	0,4999929	4,75	0,49999898	5,16	0,49999988
3,94	0,499959	4,35	0,4999932	4,76	0,49999903	5,17	0,49999988
3,95	0,499961	4,36	0,4999935	4,77	0,49999908	5,18	0,49999989
3,96	0,499963	4,37	0,4999938	4,78	0,49999912	5,19	0,49999989
3,97	0,499964	4,38	0,4999941	4,79	0,49999917	5,20	0,49999990
3,98	0,499966	4,39	0,4999943	4,80	0,49999921	5,21	0,49999991
3,99	0,499967	4,40	0,4999946	4,81	0,49999925	5,22	0,49999991
4,00	0,499968	4,41	0,4999948	4,82	0,49999928	5,23	0,49999992
4,01	0,499970	4,42	0,4999951	4,83	0,49999932	5,24	0,49999992
4,02	0,499971	4,43	0,4999953	4,84	0,49999935	5,25	0,49999992
4,03	0,499972	4,44	0,4999955	4,85	0,49999938	5,26	0,49999993
4,04	0,499973	4,45	0,4999957	4,86	0,49999941	5,27	0,49999993
4,05	0,499974	4,46	0,4999959	4,87	0,49999944	5,28	0,49999994
4,06	0,499975	4,47	0,4999961	4,88	0,49999947	5,29	0,49999994
4,07	0,499976	4,48	0,4999963	4,89	0,49999950	5,30	0,49999994
4,08	0,499977	4,49	0,4999964	4,90	0,49999952	5,31	0,49999995
4,09	0,499978	4,50	0,4999966	4,91	0,49999954	5,32	0,49999995
4,10	0,499979	4,51	0,4999968	4,92	0,49999957	5,33	0,49999995
4,11	0,499980	4,52	0,4999969	4,93	0,49999959	5,34	0,49999995
4,12	0,499981	4,53	0,4999971	4,94	0,49999961	5,35	0,49999996
4,13	0,499982	4,54	0,4999972	4,95	0,49999963	5,36	0,49999996
4,14	0,499983	4,55	0,4999973	4,96	0,49999965	5,37	0,49999996
4,15	0,499983	4,56	0,4999974	4,97	0,49999967	5,38	0,5

Приложение 2

Таблица квантилей распределения Стьюдента $t_\gamma(k)$ с k степенями свободы

$$P\left(|T| < t_\gamma(k)\right) = \gamma$$

k	γ								
	0,5	0,6	0,7	0,8	0,9	0,95	0,98	0,99	0,999
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,599
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
50	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678	3,496
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,460
80	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639	3,416
100	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626	3,390
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373

Приложение 3

Таблица значений *Z* распределения Фишера

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

r	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,010	0,020	0,030	0,040	0,050	0,060	0,070	0,080	0,090
0,1	0,100	0,110	0,121	0,131	0,141	0,151	0,161	0,172	0,182	0,192
0,2	0,203	0,213	0,224	0,234	0,245	0,255	0,266	0,277	0,288	0,299
0,3	0,310	0,321	0,332	0,343	0,354	0,365	0,377	0,388	0,400	0,412
0,4	0,424	0,436	0,448	0,460	0,472	0,485	0,497	0,510	0,523	0,536
0,5	0,549	0,563	0,576	0,590	0,604	0,618	0,633	0,648	0,662	0,678
0,6	0,693	0,709	0,725	0,741	0,758	0,775	0,793	0,811	0,829	0,848
0,7	0,867	0,887	0,908	0,929	0,950	0,973	0,996	1,020	1,045	1,071
0,8	1,099	1,127	1,157	1,188	1,221	1,256	1,293	1,333	1,376	1,422
0,9	1,472	1,528	1,589	1,658	1,738	1,832	1,946	2,092	2,298	2,647
0,99	2,647	2,700	2,759	2,826	2,903	2,994	3,106	3,250	3,453	3,800

Приложение 4

Таблица квантилей распределения Хи-квадрат $\chi^2_\gamma(k)$ с k степенями свободы

$$P\left(\chi^2 < \chi^2_\gamma(k)\right) = \gamma$$

k	γ								
	0,5	0,6	0,7	0,8	0,9	0,95	0,98	0,99	0,999
1	0,45	0,71	1,07	1,64	2,71	3,84	5,41	6,63	10,83
2	1,39	1,83	2,41	3,22	4,61	5,99	7,82	9,21	13,82
3	2,37	2,95	3,66	4,64	6,25	7,81	9,84	11,34	16,27
4	3,36	4,04	4,88	5,99	7,78	9,49	11,67	13,28	18,47
5	4,35	5,13	6,06	7,29	9,24	11,07	13,39	15,09	20,52
6	5,35	6,21	7,23	8,56	10,64	12,59	15,03	16,81	22,46
7	6,35	7,28	8,38	9,80	12,02	14,07	16,62	18,48	24,32
8	7,34	8,35	9,52	11,03	13,36	15,51	18,17	20,09	26,12
9	8,34	9,41	10,66	12,24	14,68	16,92	19,68	21,67	27,88
10	9,34	10,47	11,78	13,44	15,99	18,31	21,16	23,21	29,59
11	10,34	11,53	12,90	14,63	17,28	19,68	22,62	24,72	31,26
12	11,34	12,58	14,01	15,81	18,55	21,03	24,05	26,22	32,91
13	12,34	13,64	15,12	16,98	19,81	22,36	25,47	27,69	34,53
14	13,34	14,69	16,22	18,15	21,06	23,68	26,87	29,14	36,12
15	14,34	15,73	17,32	19,31	22,31	25,00	28,26	30,58	37,70
16	15,34	16,78	18,42	20,47	23,54	26,30	29,63	32,00	39,25
17	16,34	17,82	19,51	21,61	24,77	27,59	31,00	33,41	40,79
18	17,34	18,87	20,60	22,76	25,99	28,87	32,35	34,81	42,31
19	18,34	19,91	21,69	23,90	27,20	30,14	33,69	36,19	43,82
20	19,34	20,95	22,77	25,04	28,41	31,41	35,02	37,57	45,31
21	20,34	21,99	23,86	26,17	29,62	32,67	36,34	38,93	46,80
22	21,34	23,03	24,94	27,30	30,81	33,92	37,66	40,29	48,27
23	22,34	24,07	26,02	28,43	32,01	35,17	38,97	41,64	49,73
24	23,34	25,11	27,10	29,55	33,20	36,42	40,27	42,98	51,18
25	24,34	26,14	28,17	30,68	34,38	37,65	41,57	44,31	52,62
26	25,34	27,18	29,25	31,79	35,56	38,89	42,86	45,64	54,05
27	26,34	28,21	30,32	32,91	36,74	40,11	44,14	46,96	55,48
28	27,34	29,25	31,39	34,03	37,92	41,34	45,42	48,28	56,89
29	28,34	30,28	32,46	35,14	39,09	42,56	46,69	49,59	58,30
30	29,34	31,32	33,53	36,25	40,26	43,77	47,96	50,89	59,70
40	39,34	41,62	44,16	47,27	51,81	55,76	60,44	63,69	73,40
50	49,33	51,89	54,72	58,16	63,17	67,50	72,61	76,15	86,66
60	59,33	62,13	65,23	68,97	74,40	79,08	84,58	88,38	99,61
80	79,33	82,57	86,12	90,41	96,58	101,88	108,07	112,33	124,84
100	99,33	102,95	106,91	111,67	118,50	124,34	131,14	135,81	149,45
120	119,33	123,29	127,62	132,81	140,23	146,57	153,92	158,95	173,62

Учебное издание

Гребенникова Ирина Владимировна

МЕТОДЫ МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ
ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Редактор И. В. Коршунова
Компьютерный набор И. В. Гребенникова
Верстка О. П. Игнатьевой

Подписано в печать 25.05.2015. Формат 70×100 1/16.
Бумага писчая. Плоская печать. Гарнитура Newton.
Уч.-изд. л. 7,0. Усл. печ. л. 10,0. Тираж 100 экз.
Заказ 40.

Издательство Уральского университета
Редакционно-издательский отдел ИПЦ УрФУ
620049, Екатеринбург, ул. С. Ковалевской, 5
Тел.: 8 (343)375-48-25, 375-46-85, 374-19-41
E-mail: rio@urfu.ru

Отпечатано в Издательско-полиграфическом центре УрФУ
620075, Екатеринбург, ул. Тургенева, 4
Тел.: 8 (343) 350-56-64, 350-90-13
Факс: 8 (343) 358-93-06
E-mail: press-urfu@mail.ru

ГРЕБЕННИКОВА ИРИНА ВЛАДИМИРОВНА

Старший преподаватель кафедры информационных систем и технологий Уральского федерального университета. Окончила математико-механический факультет Уральского государственного университета им. А. М. Горького (кафедра информатики и процессов управления). Научно-педагогический стаж 14 лет. Автор 30 научных и учебных изданий. Область научных интересов — математическая теория управляемых динамических процессов: оптимизация сингулярно возмущенных систем управления с запаздыванием.